# Epigenomic studies of twins for cancer and cancer risk factors

Roos, Leonie

*Awarding institution:*
King's College London

# Epigenomic studies of twins for cancer and cancer risk factors

**Leonie Roos, M.Sc.**

A thesis submitted for the degree of

Doctor of Philosophy

Faculty of Life Sciences and Medicine

King's College London

United Kingdom

April, 2017

# Abstract

A complex interaction of environmental factors, stochastic events, and genetic susceptibility can lead to cancer development. The aim of this thesis is to investigate the DNA methylome for cancer, cancer risk, and prediction potential. Studies were performed in peripheral blood to explore systemic changes associated with cancer, and in skin for an in-depth view of total body naevus count, the strongest risk factor for melanoma.

Peripheral blood DNA methylomes of 41 cancer-discordant female monozygotic (MZ) twin-pairs were assessed for changes associated with any cancers. The epigenome-wide association study (EWAS) identified one genome-wide significant and several suggestive differential methylated positions (DMPs), three of these showed predictive biomarker potential (near *SASH1*, *COL11A2*, and *LINC00340*).

Early breast cancer specific DNA methylation changes were identified in peripheral blood obtained prior to diagnosis. The DNA methylomes were assessed by two genome-wide DNA methylation techniques in a total of 28 breast cancer discordant MZ twin-pairs. Three novel significant breast-cancer differential methylated regions (DMRs) were identified (in *MECOM*, *PCGF3*, and near *ELN*) that were suggestive of predictive biomarker potential.

Skin DNA methylomes were investigated in association with the number of naevi across the body in 322 female individuals. Three genome-wide significant DMPs were identified in novel genes *METRNL*, *C15orf48*, and *ARRDC1*. Suggestive results included *CTC1* and *RAF*, which are known genes involved in naevi

predisposition and melanoma progression. Approximately half of the 48 suggestive DMRs were correlated with gene expression in *cis*.

Overall, DNA methylation changes related to cancer, pan-cancer and breast cancer specifically, as well as with the melanoma risk factor naevus count were identified. These loci are excellent candidates for further research into their potential as biomarkers or risk factor biological mechanisms in cancer.

# Declaration

I hereby declare that this submission is my own work including all the analyses performed. To the best of my knowledge it contains no material previously published or written by another person nor material which has been accepted for any other degree of any university or other institute of higher learning, except where due acknowledgement is made in the text.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

I would like to thank my supervisor, Dr. Jordana Bell, for giving me this opportunity. You have allowed me space and freedom to continue to grow with your guidance during these years, many thanks. I would also like to say thank you to my supervisor, Prof. Tim Spector, for his valuable input in my projects and manuscripts.

A big thank you to Dr. Christopher Bell, who not only guided and encouraged me in writing my first paper during my first months, but did this throughout my PhD. Thank you for always making time to discuss my ideas and help me shape them into what is in this thesis. I truly appreciate you leaving the "me" in my writing and work.

I am most grateful to Dr. Jenny van Dongen and Prof. Dorret Boomsma from the Department of Biological Psychology, VU University Amsterdam. You have given me a great opportunity to be in your group for a month and let me delve into the great resource you have there. I enjoyed the stimulating conversation and the hospitality of the complete group.

I also would like to thank the various members of the Cancer Epigenetics group at IDIBELL Barcelona: Dr. Ana Belen, Dr. Antonio Gómez Moruno, Sebastian Moran, Dr. Holger Heyn, and Prof. Manel Esteller. It was a great experience being in your group for a good few weeks early in my PhD. In particular I would like to thank Sebastian Moran who ran the DNA methylation chips that were used in one of my projects.

To all the wonderful people at the department, thanks for all the laughs and great conversations. In particular my fellow PhD students Idil and Pei-Chien, who really helped me get started with (and love) bioinformatics during my months as a research assistant. This is why I decided to pursue this PhD, thank you.

A very special thanks to the lovely Cristina; for all the coffees, lunches at the garden museum, drinks, and all the rest. Thanks for listening to my fears, rants, and 'I can't do this's and being there for me, it means so much to me.

Margie, thank you so so much for your friendship and your continuous support and encouragement during these years. You were always there and it means so much to me. You seem to have this rare gift that when I am around you, my stress just evaporates for the moment. Thanks for all the fun times so far, and here's to way more.

To my family, jullie hebben me altijd achter me gestaan om mijn interesses to volgen overal in de wereld en dat betekent enorm veel voor me. Super bedankt voor jullie onvorwaardelijke liefde en steun.

Last but certainly not least, a big heartfelt thanks to Chris. I am not sure if I would've done this, or this much, without your constant positivity, support, and love. You are the best.

# Abbreviations

**27k** Infinium HumanMethylation27 BeadChip ..... 34

**3C** chromosome conformation capture ..... 169

**450k** Infinium HumanMethylation450 BeadChip ..... 34

**5caC** 5-carboxylcytosine ..... 27

**5fC** 5-formylcytosine ..... 27

**5hmC** 5-hydroxymethylcytosine ..... 27

**5mC** 5-methylcytosine ..... 25

**bc-DMR** breast cancer associated DMR ..... 94

**bc-DMP** breast cancer associated DMP ..... 94

**BMI** body mass index ..... 33

**BMIQ** beta-mixture quantile dilation ..... 52

**bp** base pair ..... 27

**CGI** CpG island ..... 30

**ChromHMM** chromatin state segmentation by hidden markov model ..... 28

**CNV** copy number variation ..... 114

**CpG** cytosine followed by a guanine in 5' to 3' direction ..... 25

**DISCOTWIN** discordant twin consortium ..... 39

**DMP** differentially methylated position ..... 33

**DMR** differentially methylated region ..... 31

**DNA** deoxyribonucleic acid ..... 25

**DNase-I** deoxyribonuclease I ..... 66

**DNMT** DNA methyltransferase ..... 27

**DTR** Department of Twin Research and Genetic Epidemiology ..... 43

**DZ** dizygotic ..... 37

**EBV** Epstein-Barr virus ..... 54

**ENCODE** encyclopedia of DNA elements ..... 28

**EPIC** Infinium MethylationEPIC BeadChip ..... 34

**eQTL** expression quantitative trait locus ..... 151

**ESC** embryonic stem cell ..... 25

**EWAS** epigenome-wide association study ..... 33

**FDR** false discovery rate ..... 64

**FWER** family-wise error rate ..... 65

**GEMMA** genome-wide efficient mixed model association ..... 136

**GEO** gene expression omnibus ..... 133

**GM12878** B-lymphocyte cell line transformed by Epstein-Barr Virus ..... 81

**MeDIP** methylated DNA immunoprecipitation ..... 33

**MeDIP-seq** methylated DNA immunoprecipitation sequencing ..... 34

**MIRA-seq** methylated-CpG island recovery assay sequencing ..... 147

**mQTL** methylation quantitative trait locus ..... 89

**MRE-seq** methylation-sensitive restriction enzyme sequencing ..... 168

**MRSE** methylation-sensitive restriction enzyme ..... 34

**MuTHER** multiple tissue human expression resource ..... 43

**MZ** monozygotic ..... 37

**n-DMR** total body naevus count associated DMR ..... 128

**n-DMP** total body naevus count associated DMP ..... 128

**ncRNA** non-coding RNA ..... 25

**NGS** next (second) generation sequencing ..... 33

**NHEK** epidermal keratinocytes cell line ..... 81

**NHLF** lung fibroblast cell line ..... 120

**NHS** National Health Services ..... 44

**NKR** Netherlands Cancer Registry ..... 67

**NTR** Netherlands Twin Registry ..... 67

**ONS** Office for National Statistics ..... 44

**oxBS-seq** oxidative bisulphite sequencing ..... 35

**PC** principal component ..... 52

**pc-DMR** pan-cancer associated DMR ..... 59

**pc-DMP** pan-cancer associated DMP ..... 59

**PCA** principal component analysis ..... 52

**PCR** polymerase chain reaction ..... 33

**PRC** polycomb repressive complex ..... 120

**QC** quality control ..... 48

$\mathbf{r_S}$ Spearman's rank correlation coefficient ..... 69

**RefSeq** reference sequence ..... 47

**RNA** ribonucleic acid ..... 55

**RPM** reads per million ..... 53

**RRBS** reduced representation bisulphite sequencing ..... 34

**SNP** single nucleotide polymorphism ..... 49

**SVA** surrogate variable analysis ..... 64

**TET** ten-eleven translocation methylcytosine dioxygenase ..... 27

**TFBS** transcription factor binding site ..... 66

**TSS** transcription start site ..... 30

**TwinsUK** United Kingdom adult twin registry ..... 43

**U** Unmethylated cytosine ..... 47

**UCSC** University of California, Santa Cruz ..... 66

**UK** United Kingdom ..... 43

# Publications Arising From This Thesis

### Published Journal Articles

Roos L, van Dongen J, Bell CG, Burri A, Deloukas P, Boomsma DI, Spector TD, and Bell JT. Integrative DNA methylome analysis of pan-cancer biomarkers in cancer discordant monozygotic twin-pairs. *Clinical Epigenetics* **8**, 7. ISSN:1868-7083 (2016)

Roos L, Sandling JK, Bell CG, Glass D, Mangino M, Spector TD, Deloukas P, Bataille V, and Bell JT. Higher nevus count exhibits a distinct DNA methylation signature in healthy human skin: implications for melanoma. *The Journal of Investigative Dermatology* **137**, 4, 910-920. ISSN:0022-202X (2017).

### Published Review Articles

Roos L, Spector TD, and Bell CG. Using epigenomic studies in monozygotic twins to improve our understanding of cancer. *Epigenomics* **6**, 299–309. ISSN: 1750-1911 (2014).

# Chapter 1

# Introduction

## 1.1 Epigenetics

The term epigenetics refers to mitotically and/or meiotically heritable chemical modifications of the genome that occur without any underlying change in DNA sequence, which establish and maintain cellular identity [1]. These are the mechanisms by which a cell distinguishes itself from another by specific gene expression profiles whilst having the same genome within all cells of an individual, barring somatic mutations [2]. Epigenetic mechanisms generally include molecular modifications to the DNA nucleotides themselves [3] and modifications affecting the packaging and folding of DNA around proteins that form chromatin [4, 5] (see Figure 1.1). Some consider various classes of non-coding RNAs (ncRNAs) an additional epigenetic layer [6].

### 1.1.1 DNA Modifications

To date, various covalent chemical modifications of DNA have been identified. The most well-studied and abundant mark is termed DNA methylation and comprises methylation at the carbon 5 position of cytosine (5-methylcytosine

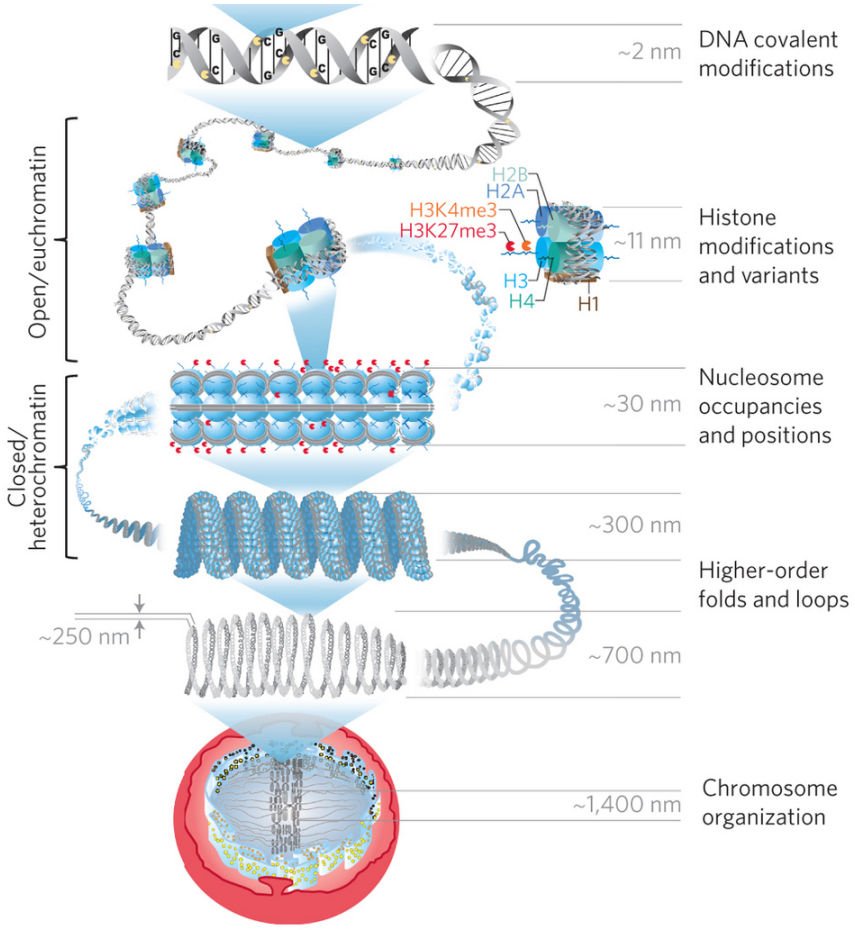



**Figure 1.1: Overview of epigenetic layers.** The base layer is seen at the top of the figure comprising the various DNA modifications. The DNA is wrapped around a histone octamer with two copies of each of the four core histones: H2A, H2B, H3, and H4 and is locked in place by linker histone H1. This allows the formation of higher order structure chromatin. The core histones can be exchanged with variants and modified at the protruding tails resulting in a dynamic structure. Reproduced and adjusted from Aguilar and Craighead [7] with permission of Nature Publishing Group.

(5mC)). The modification has been unofficially labelled as the "fifth base" of the human genome and occurs predominantly at a cytosine followed by a guanine in 5' to 3' direction (CpG) dinucleotides. This preference for a palindromic sequence enables propagation of this mark through cell division by enzymes recognising and methylating, replicated hemimethylated DNA. Although rare and not as well understood, non-CpG methylation has been found at cytosines followed by nucleotides

other than guanine [8] with prominent examples in the human brain [9] and human embryonic stem cells (ESCs) [10].

Recently, additional modifications of DNA have been identified that are products of the active demethylation pathways of 5mC. 5mC is established and maintained by DNA methyltransferases (DNMTs) and can be demethylated by passive and active pathways [11]. Passive demethylation occurs when 5mC is not faithfully maintained at replication and therefore is absent in the newly synthesised strand. Active demethylation can be catalyzed by ten-eleven translocation methylcytosine dioxygenase (TET) enzymes that produce 5-hydroxymethylcytosine (5hmC) as the first step [12, 13]. These can then be further oxidized to 5-formylcytosine (5fC) [14] and again to 5-carboxylcytosine (5caC) [15]. All of these modifications might also have distinct biological functions of their own. The most studied is 5hmC that is found to be substantially enriched in neurons of the central nervous system [16], where it could play a role via methyl CpG-binding protein 2 (MeCP2) that can bind to 5hmC [17].

### 1.1.2 Chromatin Modifications

Chromatin comprises DNA and histone proteins and provides a framework for packaging the genome within the nucleus. The fundamental unit is the multifaceted and highly dynamic nucleosome protein complex. It consists of 147 base pair (bp) of DNA wrapped in ~1.65 superhelical turns around a core histone octamer comprising two molecules of each of the four core histones: H2A, H2B, H3, and H4 [18]. Linker histone H1 binds at DNA entering and exiting the nucleosome thereby "sealing" it in place. The DNA between each pair of nucleosomes is consequently called linker DNA [19, 20]. This allows the formation of higher order structure chromatin and represents a dynamic structure that regulates access to DNA and reflects regulatory cues [4, 21]. The majority of histone modifications are at the protruding tails, nevertheless a number have been recorded at the hi-

stone globular domains interacting with other histones or DNA [22] (see Figure 1.2).

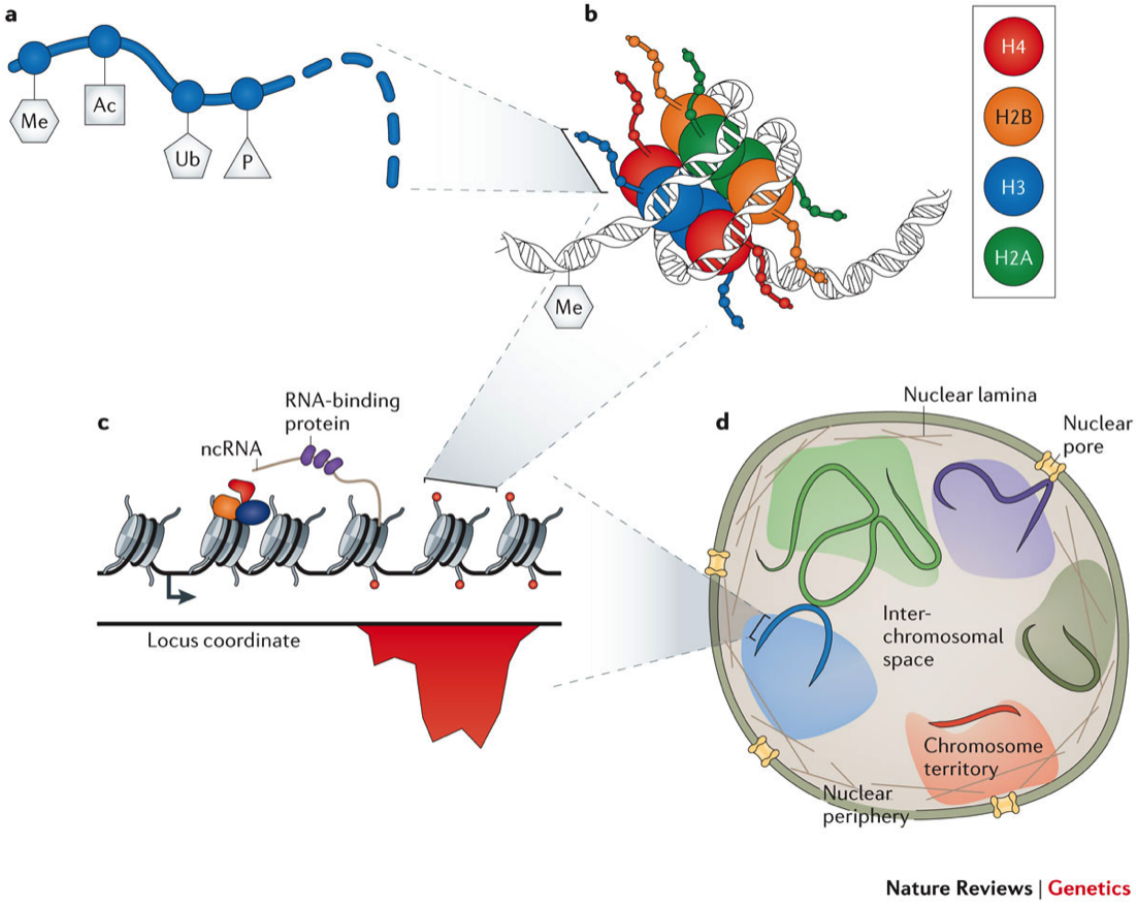



**Figure 1.2: Overview of chromatin modifications.** **(a)** Modifications of histone tails such as by the addition of methyl (Me), acetyl (Ac), ubiquitin (Ub) and phosphate (P) groups. **(b)** The nucleosome, 4 pairs of histone proteins wound by DNA that can be methylated at the cytosines. **(c)** The positioning of nucleosomes on the genome is dynamic and influences the accessibility of transcription factors to regions. Regulatory proteins (orange, blue, red and purple) can bind to nucleosomes, DNA and transcribed ncRNA. **(d)** Due to spatial configuration in the nucleus, there are interactions within and between chromosomes, as well as between with nuclear structures. Reproduced from Keung *et al.* [23] with permission of Nature Publishing Group.

Posttranslational modifications of the protruding histone tails are a key player in nucleosome dynamics and thus the organisation and function of chromatin [24]. The tails can be covalently modified at several places by numerous mechanisms

such as acetylation, methylation, and phosphorylation [25]. Thus far, research has highlighted the importance of modifications of several lysines (K) on the H3 tail. These associate with various functional elements in the mammalian genome. For example, histone H3 lysine 4 tri-methylation (H3K4me3) for active promoters, H3K4me1 and H3K27 acetylation (H3K27ac) for active enhancers, and H3K9me3 and H3K27me3 for constitutive and facultative heterochromatin respectively. Chromatin marks have been summarised into functional "segments" using the primary cell lines from the ENCODE project via a chromatin state segmentation by hidden markov model (ChromHMM) and Seqway algorithms [21, 26]. These include states annotated for example as active, weak, strong, poised, and repressed chromatin.

## 1.2 The Epigenome

The collective of the epigenetic patterns across the genome is called the epigenome. Compared to a static DNA sequence throughout life, the epigenome represents a dynamic landscape. It is essential for development where it undergoes specific changes at various stages in tidal-like waves to reset the epigenome [27–30] and establishes and maintains tissue-specific cell states [2]. The epigenome shows a continuum of change as cells mature [31]. Furthermore, throughout the lifetime of an individual, modifications at identified ageing-related loci and random "drift" have also been identified [32–36].

These epigenome changes throughout life are believed to be not only due to intrinsic and stochastic factors, but also environmental effects [37]. The epigenome can be influenced for example by prenatal and early postnatal environmental influences, such as maternal behaviour in rodents [38] and maternal diet in humans [39, 40]. Its plasticity and its essential role in gene expression make it a prime can-

didate to link environmental influences and changes in gene activity directly or in combination with genetic risk factors, influencing penetrance and expressivity.

### 1.2.1 The DNA Methylome

The DNA methylome is the complete DNA methylation pattern in the genome. In the human genome, the majority of CpGs are randomly dispersed and between 70 to 80% are methylated [41]. The frequency of CpGs is much lower than expected (at ~20% of its expected frequency) due to a high mutation rate for 5mC [42]. A small proportion of CpGs ( ~7%) cluster at higher than expected frequency in genomic regions of approximately 1 kilo base pairs (kb) known as CpG islands (CGIs) [43]. CGIs predominantly occur at gene promoters, near transcription start sites (TSSs), and first exons. They are thought to maintain their CpG content because they generally remain unmethylated in the germline [44]. The remainder of CGIs are split between intragenic and intergenic regions of the genome with the potential to act as alternate isoform promoters [45–47].

Derived from this island terminology are CGI "shores" that are the 2 kb flanking regions surrounding the CGI. These represent the boundaries of the CGIs, which still have a higher CpG frequency than expected. Shelves are the 2 kb down or upstream of the CGI shores and lead to "open sea" regions beyond. Generally, the DNA methylome in healthy cells is relatively static and faithfully maintained through cell division [48].

#### 1.2.1.1 Defining Genomic Features

The DNA methylome behaves differently at defined genomic features based on CpG density. A few genomic features of CpG density and level of DNA methylation are highlighted in this section.

### CpG Islands

Approximately 70% of genes in the human genome have CGIs in their promoter region and the majority remain unmethylated [49]. CGI methylation is typically correlated with stable epigenetically repressed regions and is associated with a range of biological processes such as genomic imprinting, X inactivation, and suppression of transposable elements [1, 3]. Methylation of promoter CGIs can be associated with stable repression of gene transcription, often at early developmental genes [50]. In general, promoter CGIs of inactive genes do not typically acquire DNA methylation but acquire tri-methatylation at H3K27 as a repressive mark [51] (see Figure 1.3). As a result, less variability in DNA methylation is observed at promoter CGIs than previously expected. This shifted the focus recently towards CGI shores that show higher variability.

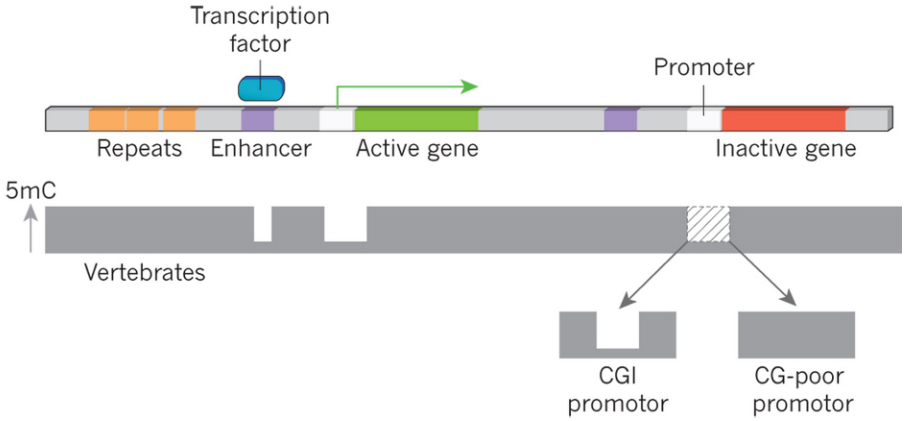



**Figure 1.3: Genomic representative region of active and inactive genes.** The region shows an example of regulatory regions including a distal enhancer and proximal promoters. The height of the bar represents the proportion of DNA methylation for the region. Modified and reproduced from Schübeler [51] with permission of Nature Publishing Group.

### CpG Islands Shores

CGI shores are enriched for differentially methylated regions (DMRs) associated with cancer-, tissue-, and reprogramming-specific changes [52–54]. Irizarry *et al.*

[52] showed that in cancer, it was the shore regions that underwent significant cancer-specific changes in DNA methylation, more so than CGIs themselves. This highlighted a loss of strict boundaries at the flanking regions of CGIs that are important events in cancer pathogenesis [52, 53]. CGI Shores also experience a tidal-like change narrowing the regions in a lineage-specific manner in tissue differentiation [54, 55] and these DMRs define tissue specificity better than changes solely based on CGIs [56].

**Low Methylation Regions**

Throughout the genome CpGs in CpG poor regions, or "open sea", are typically methylated. However, intermediate low levels of methylation with a mean of 30% are also observed and have been mapped to specific genomic regions termed low methylation regions (LMRs). These form a class of regulatory regions distal to promoters with a moderate CpG frequency. LMRs are accompanied by enhancer chromatin markers, such as H3K4me1, and act as enhancers for expression of associated promoters. They are occupied by transcription factors and their presence is required and sufficient to establish the low methylation state. Interestingly, LMRs are also more dynamic in cell differentiation than CGIs [57].

## 1.3 Epigenome-wide Association Studies (EWAS)

Complex common diseases arise from the interplay of both genetic susceptibility and environmental factors with a risk that typically increases with age [58]. This interplay between gene activity, environmental exposures, and genetics is not fully understood and the epigenomic landscape may provide critical information on the mechanisms of how environmental exposures influence penetrance and expressivity of genes [59]. Epigenetic changes independent of genetic variation, or "pure" changes, can influence disease risk, whilst those influenced by genetic variation could in theory also contribute to the “missing heritability” that is observed in

many common complex diseases [60–62].

In recent years, larger population-based epigenome-wide association studies (EWASs) have aimed to identify epigenetic differences at genomic regions associated with a phenotype of interest. The DNA methylome is the most studied amongst these and will be the focus here. Variations in the DNA methylome, or differentially methylated positions (DMPs) and DMRs, are observed in a range of complex diseases such as rheumatoid arthritis in whole blood [63], cancer in tumour tissues [64], and in multiple sclerosis in brain tissue [65]. EWASs have also identified DNA methylation variation associated with body mass index (BMI) [66], pain [67], and smoking both in adults [68, 69] and in individuals born from mothers who smoked during pregnancy [70]. In this section, DNA methylation interrogation methods will be discussed first, followed by considerations for these studies that are markedly different compared to genome-wide association studies (GWASs).

### 1.3.1 DNA Methylome Profiling Methods

Techniques for profiling of the DNA methylome are available that interrogate the DNA methylation state at a single base or regional level. Currently, three main methods are used to capture 5mC: sodium bisulphite treatment, affinity enrichment, and restriction enzymes [71]. All of these can be followed by high-throughput array analysis or next (second) generation sequencing (NGS) (see Figure 1.4).

#### 1.3.1.1 5mC Assessment

Sodium bisulphite chemically deaminates unmethylated cytosines to uracil in denatured genomic DNA, whilst 5mC remains unaffected. When amplified using polymerase chain reaction (PCR), the uracils are converted to thymines. This process thus converts an epigenetic difference into a genetic difference, thereby enabling detection.

Affinity enrichment uses either antibodies for 5mC or specific methyl

binders such as methyl binding domain (MBD) proteins to bind methylated fragmented (denatured) DNA. Antibodies that bind directly to methylated DNA and are followed by immunoprecipitation are used for the methylated DNA immunoprecipitation (MeDIP) strategy [72]. In contrast, in a MBD-based approach the antibodies bind to the methyl binding protein [73].

Restriction endonuclease methods rely on methylation-sensitive restriction enzymes (MRSEs) combined with methylation-insensitive restriction enzymes to assess the DNA methylation status. MRSEs will generate fragments dependent on methylation status that is compared to fragments of an isoschizomer that is unhindered by methylation status for the same recognition site. Thereby it obtains information about the methylation status at a single base level. A widely used pair of restriction enzymes are the methylation sensitive HpaII and insensitive MspI [74].

#### 1.3.1.2 Comparison of Methods

Each method has its advantages and disadvantages with considerations depending on coverage of the genome, costs, analysis methodology, and bias. For example, array-based methods offer a cost-effective approach, but are limited in the amount of CpGs across the genome and typically biased towards gene rich regions. NGS based approaches offer the advantage of increased coverage, but are more costly and can be biased towards CpG dense regions, with the exception of whole genome bisulphite sequencing (WGBS).

To date, the most used technique is bisulphite conversion followed by array-based analysis covering ~480,000 CpGs, the Infinium HumanMethylation450 BeadChip (450k), and its predecessor the Infinium HumanMethylation27 BeadChip (27k) covering ~27,000 CpGs. The 450k was discontinued early 2016 and is now replaced by the Infinium MethylationEPIC BeadChip (EPIC) covering ~850,000 CpGs [75].

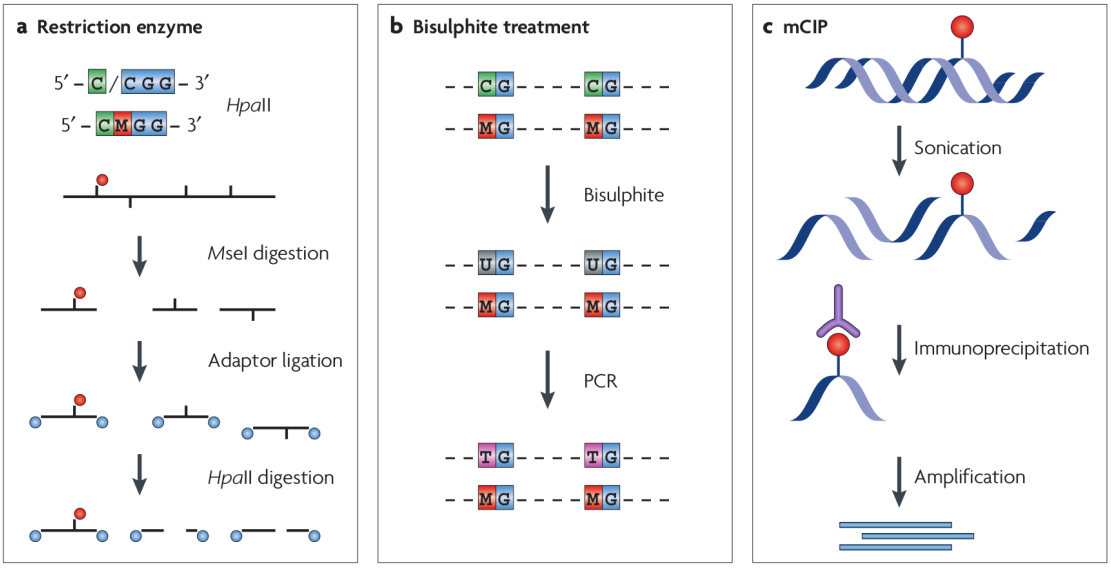



**Figure 1.4: Assay methods for 5mC. (a)** Restriction enzymes showing the methylation sensitive HpaII and insensitive MspI. **(b)** Bisulphite treatment that converts unmethylated cytosines to uracils whilst leaving methylated cytosines unchanged. **(c)** Immunoprecipitation of methylated DNA (here mCIP). An antibody is used on sonicated DNA to pull down methylated regions. All of these methods can be followed with microarrays or sequencing. Modified and reproduced from Schones and Zhao [71] with permission of Nature Publishing Group.

Sequencing based approaches include the "gold standard", WGBS, that interrogates every single base in the genome. This approach is the most expensive and includes sequencing 70 to 80% of genome that does not contain variable CpGs dinucleotides [41]. Methods that reduce the amount of starting material for NGS and thus reduce costs, include methylated DNA immunoprecipitation sequencing (MeDIP-seq) and reduced representation bisulphite sequencing (RRBS). MeDIP-seq uses an antibody for 5mC to only capture methylated DNA fragments for sequencing and provides regional information. RRBS uses restriction enzymes to generate genomic DNA fragments followed by bisulphite sodium treatment for sequencing and provides a single base resolution.

The bisulphite conversion reaction can not distinguish 5mC from oxidised variants, such as 5hmC. However, the abundance of 5hmC in the human genome is relatively low and is more prevalent in certain tissues [76]. In recent years further

methods have become available for identifying oxidised variants like 5hmC, such as oxidative bisulphite sequencing (oxBS-seq) [77].

### 1.3.2 Considerations for EWAS

There are a number of considerations for EWAS designs due the dynamic nature of the epigenome, cell specificity, confounders, and interpretation of results. The DNA methylome varies throughout the lifetime of an individual, across different cell types, lineages, and maturation states. Apart from recently developed single-cell methods [78], the vast majority of DNA methylomes is profiled using a sample of heterogeneous/mixed cells, resulting in a relative profile reflecting the composition of the sample. The field of epigenomic epidemiology is rapidly refining its research questions on these observed dynamics.

A direct result of the increasing knowledge of DNA methylomes across vast numbers of different tissues and cell lines [2], is the importance of homogeneity in samples. Most studies are based on whole blood samples owing to ease of accessibility and the premise that a blood sample can reflect the health of an individual. A blood sample however, is heterogeneous in its composition and therefore heterogeneous in the DNA methylomes present. Preferably, cellular heterogeneity should be taken into account as composition between samples can lead to confounding such as observed with inflammation and smoking for whole blood samples for example [63, 79]. EWASs of a heterogeneous samples can use analytical frameworks that can estimate proportions of cell types such as the Houseman algorithm for whole blood samples assessed on array based platforms [80]. However, this should not be standard practice if diseases or phenotypes are associated with differences in composition. The interpretation of results and its biological value should take this into account. A hypothesis that investigates an in-depth biological view of a phenotype, should ideally profile the DNA methylome in the cell type or tissue that is of greatest biological relevance to the phenotype.

In contrast with GWAS, EWAS results are not necessarily causal to the phenotype. Epigenomic variation can indeed contribute to the pathogenesis (causal), be a consequence of it (reverse causality), or a confounding variable [81]. When interpreting the results, this is crucial to address if any inference can be drawn from the results. With longitudinal cohorts this can be achieved by using samples that are sampled before disease onset to determine if the changes precede it. Alternatively, Mendelian-randomization using a genetic proxy or causal-inference tests can be performed to test causal dependency [81, 82]. However, even in the absence of such longitudinal data and causal inference analyses, these EWAS signals can still be of interest in terms of their potential as biomarkers.

Each method of analysing the DNA methylome can have their own bias and the data can be affected by technical confounders such as batch, plate, and/or array that should be addressed in the downstream analyses. A randomised design is also preferably implemented before any downstream profiling or analysis. Apart from technical confounders, there are also biological confounders for DNA methylome data such as age, gender, and genetic variation. These should ideally be taken into account at the design stage as well as in the analysis.

Factors that can influence the power of an EWAS include effect size, number of individuals, and the multiple testing threshold for significance. These and the aforementioned epidemiological considerations should all be taken into account to optimise power to detect significant phenotype-associated epigenetic variation.

## 1.4 The Twin Model

For decades, classical twin studies have greatly contributed to complex trait variation by disentangling genetic and environmental influences [83]. Human identical, or monozygotic (MZ), twins arise from the separation of daughter cells of a single fertilized ovum and thus have close to identical genomes. Non-identical, or

dizygotic (DZ), twins arise from two independently fertilised ova and share on average 50% of segregating DNA sequence variation similar to normal siblings [84, 85]. Both MZ and DZ twin-pairs share early environment however, DZ twin-pairs have a much-reduced genetic component that forms the basis of the classical twin design. MZ twin-pairs have surprisingly high discordance rates for common complex disease including metabolic disease, autoimmune disease, and cancer [86–90]. This difference is typically attributed to environmental influences, although this component will include stochastic effects as well [91].

Epigenomic variation can be quantified and associated with discordance between MZ twin-pairs in complex traits. Early studies using limited number of CpGs showed that MZ twin-pairs have in fact more similar DNA methylomes than DZ twin-pairs and that differences between MZ twin-pairs increase with age [33, 34, 92, 93].

### 1.4.1 The Discordant Monozygotic Twin Design for EWAS

The discordant MZ twin model should elucidate epigenomic variation independent of genetics or other confounding factors. MZ twin-pairs are an ideally matched case-control study as they are matched for nearly all genetic factors, age, cohort effects, and common early environment [94]. The advent of high-throughput DNA methylome techniques has driven an exponential rise in discordant MZ EWASs. Differences in the DNA methylome of disease-discordant MZ twins have been identified at several new candidate genes in a range of tissues and complex diseases and traits.

The majority of discordant MZ EWASs are in whole blood samples or blood derived cells. A range of differences have been identified in blood samples for complex diseases such as type 1 and type 2 diabetes [95, 96], bipolar disorder [97], systemic lupus erythematous (SLE) [98], scleroderma [99], autism [100, 101], and schizophrenia [102]. Furthermore, DNA methylome differences associated with

traits and environmental factors have been identified such as smoking [103], obesity in subcutaneous adipose tissue [104] and leukocytes [105], birth weight [106], and pain sensitivity [67].

These studies have revealed DNA methylome changes broadly independent of genome sequence variation in a large number of regions across the genome. Larger consortia have now been set up with the purpose of collecting more samples of discordant MZ twin-pairs and for meta-analyses across cohorts such as discordant twin consortium (DISCOTWIN) [107].

## 1.5 Cancer Epidemiology

Despite great global efforts in cancer research, cancer is still one of the leading causes of death in industrialized societies, second only to cardiovascular disease [108, 109] (see Figure 1.5). It is heterogeneous and encompasses a large set of diseases all characterised by uncontrolled cell proliferation and loss of differentiation. Human tumours arise through a multistep process and are defined by six hallmarks: proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastatic pathways [110, 111]. This progression is driven by the interplay between genetic mutations and epigenetic variation [112]. Highly penetrant genetic variants have been identified in families with rare hereditary cancer syndromes that account for 5-10% of all cancers [113–116], whereas contribution of common inherited genetic variation in sporadic cancers cases is moderate. Concordance rates of cancer among MZ twin-pairs are generally below 0.15. For the commonest types, the contribution of the non-genetic component was estimated to be 60% to 70% [90]. Sporadic cancers have been associated with a number of risk factors by epidemiological studies such as age, smoking, obesity, and alcohol consumption. The epigenome is seen as the prime candidate for mediating or quantifying these

risk factors.

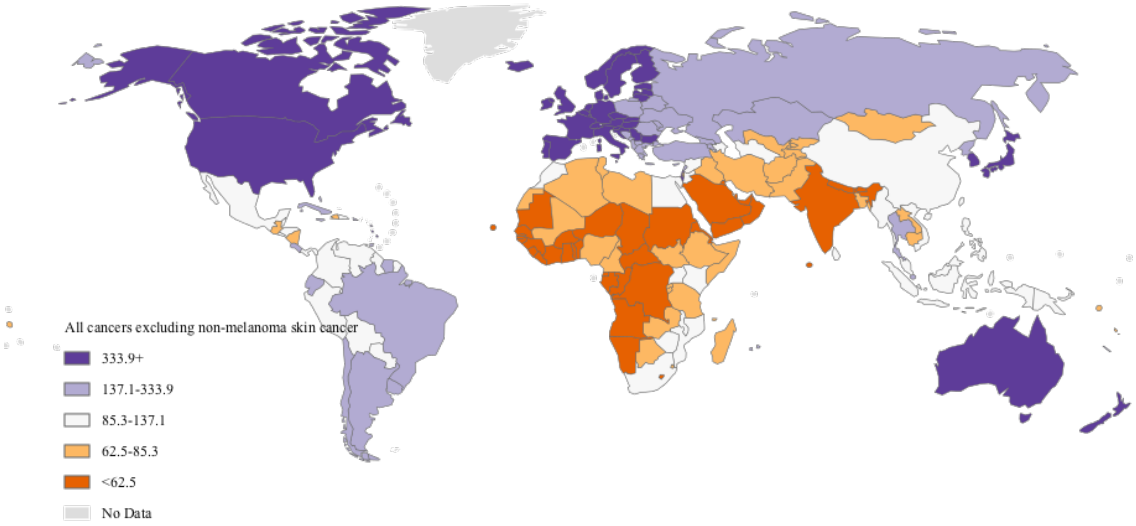



**Figure 1.5: One year cancer prevalence proportions per 100,000 individuals including both sexes.** All cancers are included except non-melanoma skin cancers data from 2012. Reproduced from data source: GLOBOCAN project, part of International Agency for Research on Cancer (IARC) [117].

### 1.5.1 DNA Methylome in Cancer

An aberrant DNA methylome, both in terms of global trends and promoter specific marks, is a hallmark of cancer cells [118, 119]. This highlights the importance of a functional DNA methylome in healthy cells. Changes in the epigenome appear to affect more genes compared to genetic mutations and plays a pivotal role in cancer [120]. This could be driven by the high level of somatic mutation in genes that read and write the epigenome [121]. A global loss of DNA methylation is observed in large blocks that cover over half of the genome that is accompanied by gene silencing. Additionally, a gain or loss of DNA methylation can occur at genes that can mimic genetic mutations [53, 122, 123]. This is all indicative of an interplay of genetic and epigenetic variation that provide a growth advantage for

cancer cells [124]. Some of these changes in DNA methylation can be observed in early neoplastic tissues [125, 126] and even in healthy tissues as risk factor related signatures [127].

### 1.5.2 DNA Methylome as Biomarker of Cancer

A complex combination of environmental factors, stochastic events, and genetic variation can lead to cancer pathogenesis. Early diagnosis greatly improves the odds of successful treatment and life expectancy. Thus, identifying individuals at risk or early stages without an invasive or costly procedure is highly desirable. One direction in cancer research is to quantify this combination and detect pre-disease changes or states to use as systemic biomarkers. The DNA methylome lends itself well for this considering its aberrant landscape in tumours itself, relative stability in various sample conditions, and previous successes with complex disease [128]. Deconvolution of plasma DNA methylomes of cancer patients also show more contribution of DNA from cells of the primary cancer location as well as tumour DNA [129]. DNA methylome markers have been established using non-invasive tissues such as circulating DNA in plasma [130], serum [131], and urine [132] as well as DNA methylation changes in sputum [133] or blood samples [134].

Blood sample tests are used to check the general health status of an individual using haematological measures. On this premise, peripheral blood or blood cell types might contain specific or systemic changes that are associated with pre-disease state, or the body's response to (metastatic) cancer. Indeed, DNA methylation signatures in peripheral blood have been not only associated with risk factors such as smoking [69] and age [135] but also with cancer development at primary locations including breast [136, 137], colon [138], bladder [139], and ovary [140].

## 1.6 Thesis Outline

The overall aim of this thesis is to gain insight into DNA methylome differences in cancer discordant MZ twin-pairs as well as an in-depth view of the strongest risk factor for melanoma. This leads to the following specific aims:

**Cancer biomarker using MZ twin-pairs**

The first aim is to examine evidence for a whole blood DNA methylation signature common to multiple cancers using the discordant MZ twin design. To address this I selected 41 MZ twin-pairs with blood samples taken within a 5 year time frame of diagnosis and also explored DNA methylation stability over time.

The second aim is to identify a blood-based DNA methylation signature specifically in breast cancer discordant MZ twin-pairs sampled up to 8 years prior to diagnosis. Here, breast cancer discordant MZ twin-pairs were selected from two epigenome-wide DNA methylation interrogation methods: 28 pairs profiled by the 450k and 26 pairs profiled by MeDIP-seq. The results were explored across platforms as well as for stability over the years preceding diagnosis.

**Integrative view of DNA methylome in skin tissue for naevus count**

The third aim was to investigate DNA methylation variation in skin tissue associated with total body naevus count, the strongest risk factor for melanoma. I assessed skin DNA methylomes in 322 individuals for naevus count and examined gene expression as well as genetic variants previously identified for naevus count and melanoma by GWAS.

# Chapter 2

# Material and Methods

## 2.1 Subjects

All subjects included in this thesis are volunteer adult twins from across the United Kingdom (UK) and registered at the United Kingdom adult twin registry (TwinsUK). The TwinsUK is hosted by the Department of Twin Research and Genetic Epidemiology (DTR), King's College London (KCL), and is based at St. Thomas' Hospital in London.

### 2.1.1 TwinsUK Cohort

The TwinsUK started in 1992 and now comprises approximately 12,000 MZ and DZ twins within the age range of 18 to 103. It is largely female ( ~80%) and is approximately evenly divided between MZs and DZs [141]. The twins in this cohort are not selected for diseases and do not differ in means and ranges of quantitative phenotypes to an age-matched population in the UK [142]. The primary research focus of TwinsUK is epidemiology studies of complex diseases and healthy ageing using multiple "omic" technologies.

#### 2.1.1.1 Biological Samples

The TwinsUK has built a substantial biobank of biological samples that are continually donated to by the twins. Biological sample collections relevant to this thesis include over 120,000 blood samples from approximately 8,000 twins [141], and ~850 twins that have provided blood samples for generating lymphoblastoid cell lines (LCLs) and undergone a punch biopsy for skin and adipose tissue under the umbrella of the multiple tissue human expression resource (MuTHER) project [143].

#### 2.1.1.2 Phenotype Collection

About half of the registered twin-pairs attended a comprehensive clinical visit with a substantial subset attending more than one follow-up visit over the last two decades. During this visit various clinical tests, such as bone mineral density scans, height, and weight, as well as biochemical assays were performed. Over 7,000 twins responded to at least one of the annual questionnaires to assess for conventional epidemiological phenotypes [141]. All twins are linked with their National Health Services (NHS) numbers to the Office for National Statistics (ONS) of England for detailed mortality information as well as detailed cancer diagnosis if made by a UK registered pathologist.

### 2.1.2 Ethical Approval

Written informed consent for multi-omic analysis from all subjects in the TwinsUK cohort was obtained in accordance with Guy's & St Thomas' NHS Foundation Trust Ethics Committee (EC04/015 - 15-Mar-04).

## 2.2 Phenotype Selection

The selected and analysed phenotypes pertaining each subset are described in detail within each chapter. The cancer classification system is described here.

### 2.2.1 Cancer Diagnosis

Detailed cancer diagnosis information used in chapter 3 and 4 was obtained through record linkage with the ONS registry. This registry includes the World Health Organization (WHO) diagnostic code system of disease classification: the International Classification of Diseases (ICD)-10 [144]. Codes ranging from C00 to C96 provide information on neoplasms. The most recent ONS record linkage is from June 2015.

## 2.3 DNA Methylome Profiling

Two different methods were used to profile the DNA methylome in this thesis. These included the widely used array-based method, 450k and a NGS based method, MeDIP-seq. Peripheral blood DNA methylomes were analysed in chapters 3 and 4, and skin DNA methylomes in chapter 5.

### 2.3.1 Infinium HumanMethylation450 BeadChip (450k)

The DNA methylomes used in this thesis were selected from three larger 450k DNA methylation datasets profiled as part of previously funded research in the TwinsUK cohort. Two of these included peripheral blood DNA methylomes that were profiled at two different genomic centres; the Wellcome Trust Sanger Institute and IDIBELL. The first comprised 957 samples of 915 unique individuals profiled at the Wellcome Trust Sanger Institute and the second comprised 104 samples of 86 unique individuals profiled at IDIBELL. Part of this second subset was previously

published in a breast cancer study [137]. Skin DNA methylomes were selected from one larger dataset comprised of dissected skin tissue from 468 punch biopsies from female twins as part of the previously published MuTHER project [145, 146].

#### 2.3.1.1 Illumina 450k Probe Design

The array methodology is based on the sodium bisulphite conversion of unmethylated cytosines to uracil and subsequently post-PCR to thymine. The array then evaluates these C/T generated polymorphisms to quantify DNA methylation with 12 samples per beadchip. The introduction of a new probe design (Infinium II) to quantify DNA methylation drove the large increase in the number of cytosine sites now measured on the 450k across the genome (485,764). However, the 450k still contains the previous probe design (Infinium I) of its predecessor, the 27k, and is a mix of these two probe designs: Infinium I ( ~30%) and Infinium II ( ~70%).

The Infinium I technology comprises two different probes on two separate beads, with the 3' end terminus of each probe designed to match either the methylated or unmethylated allele. It is followed by a single base extension right after the CpG dinucleotide, that is fluorescently labelled and measured separately for the unmethylated and methylated probe sequences. The Infinium II uses one probe type on one bead to detect methylated or unmethylated alleles. Here, the single base extension of the probe determines the methylated or unmethylated state by complementing either the C or T measured in green or red channels respectively [147].

The DNA methylation levels are represented by betas, denoting the ratio of the methylated signal over the denominator, being the sum of unmethylated and methylated signals plus a constant of 100. This results in a beta value between 0, not methylated, and 1, methylated (see Figure 2.1). Due to the biological nature of the DNA methylome, the vast majority of probes measured possess a low or high beta value and only few intermediately methylated sites. Therefore, the beta

density plots per individual typically show a bimodal distribution with two peaks close to 0 and 1. These bimodal peaks are slightly closer to 0 and 1 for Infinium I than Infinium II (see Figure 2.2). The Infinium II is less sensitive to identify extreme methylation values because of the single probe method that introduces a binding competition for unmethylated and methylated cytosines [147].

$$beta = \frac{Intensity(M)}{Intensity(M+U)+100}$$

**Figure 2.1: Beta value equation.** Methylated cytosine (M) intensities over the sum of both M and Unmethylated cytosine (U) plus a constant of 100.

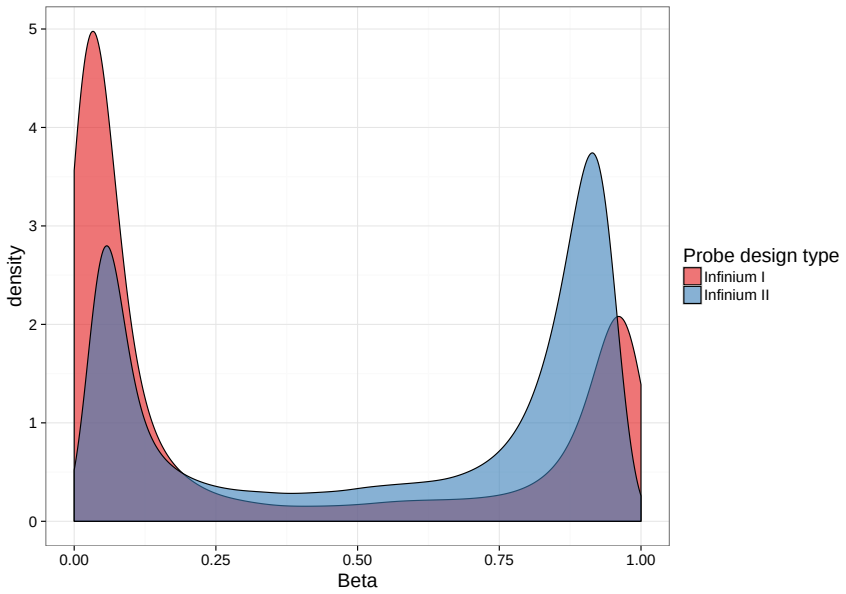



**Figure 2.2: Density distribution of one sample per Infinium design.**

#### 2.3.1.2 Distribution of CpG Sites

The 450k contains 482,421 probes that target CpG dinucleotides in the human genome distributed over all chromosomes. In terms of CpG density of the targeted CpGs, 30.9% are in CGIs, 23% in CGI shores, 9.7% in CGI shelves, and 36.3% are isolated CpGs in the "open sea" [148]. The 450k covers at least one

CpG site in 96% of all human CGIs [147]. Furthermore, 99% of reference sequence (RefSeq) genes are screened on the beadchip, with CpG sites located across gene regions in the following manner: 20.75% in promoter regions, 5% in 5' untranslated regions (UTRs), 32.30% in gene bodies, and 3% in 3' UTRs (see Figure 2.3). Approximately a quarter of the probes are located in intergenic regions [147]. A minority, 0.85%, are interrogating ncRNAs [148].

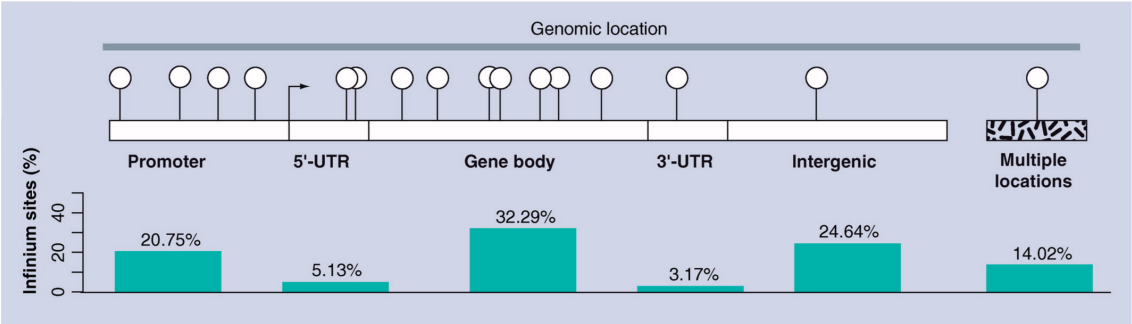



**Figure 2.3: Distribution of CpG sites on the 450k across gene regions.** Histograms showing the percentage of cytosines covered by the 450k by their genomic location. Reproduced and adjusted from Dedeurwaerder *et al.* [147] with permission of Future Medicine Ltd.

#### 2.3.1.3 Quality Control

Each of the DNA methylome datasets used in this thesis were assessed for quality control (QC) as outlined below. The QC is performed within R with use of additional packages from Bioconductor for reading the raw output file format of Illumina when available (Minfi [149]).

**Probe level QC**

The data was read into R and the following five sets of probes were removed for the main analyses that: 1) failed detection in one or more samples and/or had a bead count less than 3 in >5% of samples (n = variable for each dataset, see details in each chapter), 2) aligned to more than one location in the human genome with

their 50 bp sequence (n = 17,764), 3) located on the sex chromosomes (n = 11,650), 4) harboured common genetic variants occurring in European Caucasians (minor allele frequency (MAF) >1%) within 10 bp on the probe at the interrogated CpG site (15,827), and 5) contained variants at any MAF at the interrogated CpG site (11,236) based on data from the 1000 Genomes Project [150, 151].

**Sample level QC**

Samples were inspected visually for outliers across all CpG sites using boxplots (their mean and median DNA methylation), beta density plots, and heatmaps. Known intermediate methylated loci at established imprinted regions were also assessed [152].

#### 2.3.1.4 Sample Identification

The individuals from each DNA methylome data subset used in this thesis were verified using their known genotype available from TwinsUK single nucleotide polymorphism (SNP) array data (Illumina HumanHap300, HumanHap610Q, 1M-Duo, or 1.2M-Duo custom arrays, see Section 2.5 on page 55). To this end, a sample identifier was created and tested as described in this section.

The sample identifier uses the 57 autosomal quality control probes on the 450k that do not interrogate DNA methylation but interrogate common SNPs for identification purposes. The beta value for these probes is calculated in the same way as the rest of the beadchip and thus produces a value between 0 and 1 (see Figure 2.1 on page 47). No further information is provided for the actual genotype but it can be deducted from the equation that both values of 0 and 1 indicate a homozygous state for the two alleles, whereas heterozygous individuals will average around 0.5. These 57 SNPs are common and selected for sample identification and have a average heterogeneity in the range of 0.41-0.5 (1000 Genomes data phase 3). Consequently, the assumption was that the heterozygous state will be

present enough and over enough SNPs, to provide adequate power to build a sample identifier on only these two categories of either heterozygous or homozygous genotype states (see Figure 2.5 on page 56 for an schematic overview).

The sample identifier was first tested and refined on the complete 450k dataset profiled at the Wellcome Trust Sanger Institute. This comprised of 957 peripheral blood samples of 915 unique individuals described in section 2.3.1. Of these, 927 samples were genotyped in the TwinsUK on SNP arrays and 30 were not yet genotyped.

To evaluate the sensitivity of the sample identifier, all twins from the TwinsUK with available genotype data, ~5,700 individuals, were used. The 57 autosomal SNPs were extracted with PLINK [153] and were recoded as 0 for homozygous alleles and 1 for heterozygous alleles. For all 927 samples from the 450k, the 57 SNP control probes were selected and clustered into the three states using k-means clustering: two homozygous ( ~0.1 and ~0.9) and one heterozygous ( ~0.5) state, later recoded to 0 and 1 respectively. Subsequently, the 57 allelic states were matched per individual from the 450k to the 57 allelic states of all ~5,700 TwinsUK individuals from the SNP arrays. A match was defined per SNP as either consistent homozygous (0) or heterozygous (1), while a mismatch was defined as an inconsistent state (see Figure 2.4).

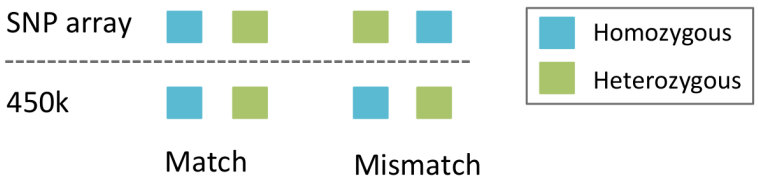



**Figure 2.4: Possible allelic combinations per SNP and match calling.**

The most number of matches ("best fit") across the 57 SNPs were recorded for all 927 samples from the 450k. This resulted in a total of 94.7% that matched

perfectly on all 57 SNPs (878 samples), *i.e* all allelic states for each SNP were similar. A further 2.6% matched with more than 54 matches for the total of 57 SNPs (24 samples). This totaled 902 samples, of which 33 samples paired to a different individual identifier. The vast majority (30) of these occurred in the first profiling batch and paired with individuals not profiled on the 450k. This is most likely due to sample mix ups described at the site. The remaining three occurred in the second profiling batch and appear swapped on one 450k beadchip.

Next, the second most number of matches for these 902 samples were examined to assess specificity. Here a minimum of 12 mismatches was observed with a mean of 15 mismatches. Thus, when there were little mismatches in the best fit (<3), the second best fit showed sufficient difference across all 57 SNPs.

Lastly, 30 samples from the 450k who were not yet genotyped were assessed to further check sensitivity as these should not match any genotype on file. These were also matched to all ~5,700 individuals as described above. Here the minimum of the least number of mismatches across the 30 samples was 14 with a mean of 27.

To conclude, a high percentage of true matches ( ~97%) was observed in for the sample of 927 with genotype data available as expected. There were no matches observed for 450k samples that were not yet genotyped. Therefore it was concluded that this method was sensitive enough to use as a sample identifier for the 450k.

#### 2.3.1.5 Peripheral Blood Cell Proportions

Cell type proportions were estimated directly on the beta values using the method devised by Houseman *et al.* [80] for six blood cell types; CD8+ T cells, CD4+ T cells, B cells, Natural Killer cells, granulocytes, and monocytes. This was done for all peripheral blood DNA methylomes and are described in more detail in the chapters pertaining to the selected subsets. Subsequently, pairwise correlations between these proportions were assessed by Spearman's Rank correlation to explore the relationships between the proportions for downstream analyses.

#### 2.3.1.6 Normalisation

The DNA methylomes were normalised using the intra-array normalisation, beta-mixture quantile dilation (BMIQ) [154], to correct for probe type bias in chapter 3 and 5. BMIQ normalisation adjusts the beta values of Infinium II design probes (70%) into the distribution characteristics of Infinium I probes. This makes effect sizes of probes across the two types more comparable. In chapter 4, the beta values were normalised using functional normalisation that removes technical variation using control probes [155]. Additionally, each probe was then standardised to N(0,1) for downstream analyses.

#### 2.3.1.7 Identifying Confounders

Per chapter, principal component analysis (PCA) was performed on standardised beta values (N(0,1) for each DNA methylome dataset. The first few principal components (PCs) of each dataset were examined for the proportion of total variance explained, and tested for associations with potential covariates for DNA methylome data such as beadchip, sample position on the beadchip, age, batch of profiling, phenotype of interest, BMI, smoking, bisulphite conversion efficiency (when available), and estimated blood cell proportions [80] for whole blood samples. The results are presented in each chapter.

### 2.3.2 Methylated DNA Immunoprecipitation Sequencing (MeDIP-seq)

The peripheral blood DNA methylomes in chapter 4 were selected from a larger dataset of ~5,000 twins (the EpiTwin project) profiled by MeDIP-seq.

#### 2.3.2.1 MeDIP-seq Design

MeDIP is an affinity enrichment that uses antibodies for 5mC to bind to methylated fragmented and denatured DNA [72]. In MeDIP-seq only these fragments (containing 5mC) are sequenced by NGS and provide regional information on DNA methylation (DMRs) [156]. Thus in theory, this provides a genome-wide coverage of potentially all methylated CpGs.

The design and workflow is shown in Figure 2.6. MeDIP-seq starts with purified DNA that is fragmented by sonication. To increase the affinity of the antibody, these fragments are denatured. The denatured fragments are then incubated with antibodies for 5mC. This is followed by immunoprecipitation with antibodies with conjugated magnetic beads against the bound 5mC antibodies. The unbound DNA is removed with the supernatant. The antibodies are then digested and the DNA fragments are used for downstream NGS (see Figure 2.6) [72, 156].

After sequencing of the methylated DNA fragments, the reads are aligned to a reference genome. For MeDIP-seq specific analysis, packages such as MEDIPS [157] can be used for QC and generation of genomic windows, or bins, of DNA methylation scores. Commonly this may be 500 bp windows with a window overlap of 250 bp. These scores include reads per million (RPM) per bin, that with standardisation enables comparability between samples.

#### 2.3.2.2 EpiTwin MeDIP-seq Dataset

The total EpiTwin MeDIP-seq dataset includes ~5,000 twins. All DNA sample preparation, MeDIP reaction, and Illumina NGS was performed at BGI,

Shenzen, China. The sample preparation and initial QC and alignment was performed by BGI in collaboration with the DTR. This is described in detail for a subset by Davies *et al.* [158] and is summarised below. Further details and QC specific for the subset assessed is detailed in chapter 4.

**Sample Preparation**

DNA was fragmented using a Covaris sonication system and libraries for sequencing were prepared from 5 µg fragmented genomic DNA. End repair, <A> base addition and adaptor ligation steps were performed using Illumina's Single-End DNA Sample Prep kit. The anti-5mC commercial antibody (Diagenode) was used to immunoprecipitate the Adaptor-ligated DNA, and the MeDIP resultant was validated by quantitative PCR. The MeDIP DNA was then purified with ZYMO DNA Clean & Concentrator-5 columns, followed by amplification using adaptor-mediated PCR. Fragments between 220 and 320 bp were selected by gel excision followed by QC evaluation by Agilent BioAnalyzer analysis [156]. These libraries were then subjected to highly parallel 50 bp single-end sequencing on the Illumina HiSeq platform.

**QC and Alignment**

Raw sequencing data passed initial QC using in-house scripts and FastQC [159]. An average of 17 million uniquely mapped reads were obtained for each of the samples. Alignment to hg19 was performed with BWA [160] and MEDIPS [157] was used to calculate RPM scores in defined bin sizes of 500 bp with an overlap of 250 bp across the genome. Further visual QC was performed in R via a correlation matrix, hierarchical clustering, dendrogram, heatmap, and density plots. Batch effects were assessed via PCA. The total number of bins was 12,382,723. This was later reduced to autosomes only for further analysis (11,524,145).

## 2.4 Gene Expression Profiling

Three tissues for transcriptome profiling were collected at the same clinical visit for 856 healthy female twins from the TwinsUK by the MuTHER project as previously described [146]. In short, punch biopsies (8mm) were done at photo-protected area adjacent and inferior to the umbilicus of which subcutaneous adipose tissue and skin tissue were dissected and stored in liquid nitrogen. LCLs were generated through Epstein-Barr virus (EBV)-mediated transformation of B-lymphocytes from whole blood samples.

RNA was extracted from the tissues and the Illumina Human HT-12 V3 BeadChips were used for transcriptomic profiling. Each sample had three technical replicates. Probes with less than 3 beads were removed and remaining expression signals were $\log_2$-transformed. They were normalised with quantile normalization of the replicates of each individual followed by quantile normalization across all individuals for each tissue.

The Illumina gene expression beadchips include 48,804 probes that target approximately 25,000 annotated genes using RefSeq and the UniGene databases. It provides a genome-wide transcriptome coverage also including splice variants.

## 2.5 SNP Genotyping

Genotyping of individuals of the TwinsUK was done using a combination of Illumina HumanHap300, HumanHap610Q, 1M-Duo, or 1.2M-Duo custom arrays. Imputation was performed with IMPUTE using the 1000 Genomes data phase 3 reference panel, as previously described [161]. Quality control genotype measures included thresholds for minimum genotyping rate (>95%), Hardy–Weinberg equilibrium ($p$ >1.0 x $10^{-6}$), and MAF (>1%). The imputation quality score was >0.8 for SNPs used in the sample identifier (described in section 2.3.1.4) and >0.5 for GWAS catalogue SNPs specifically in chapter 5.

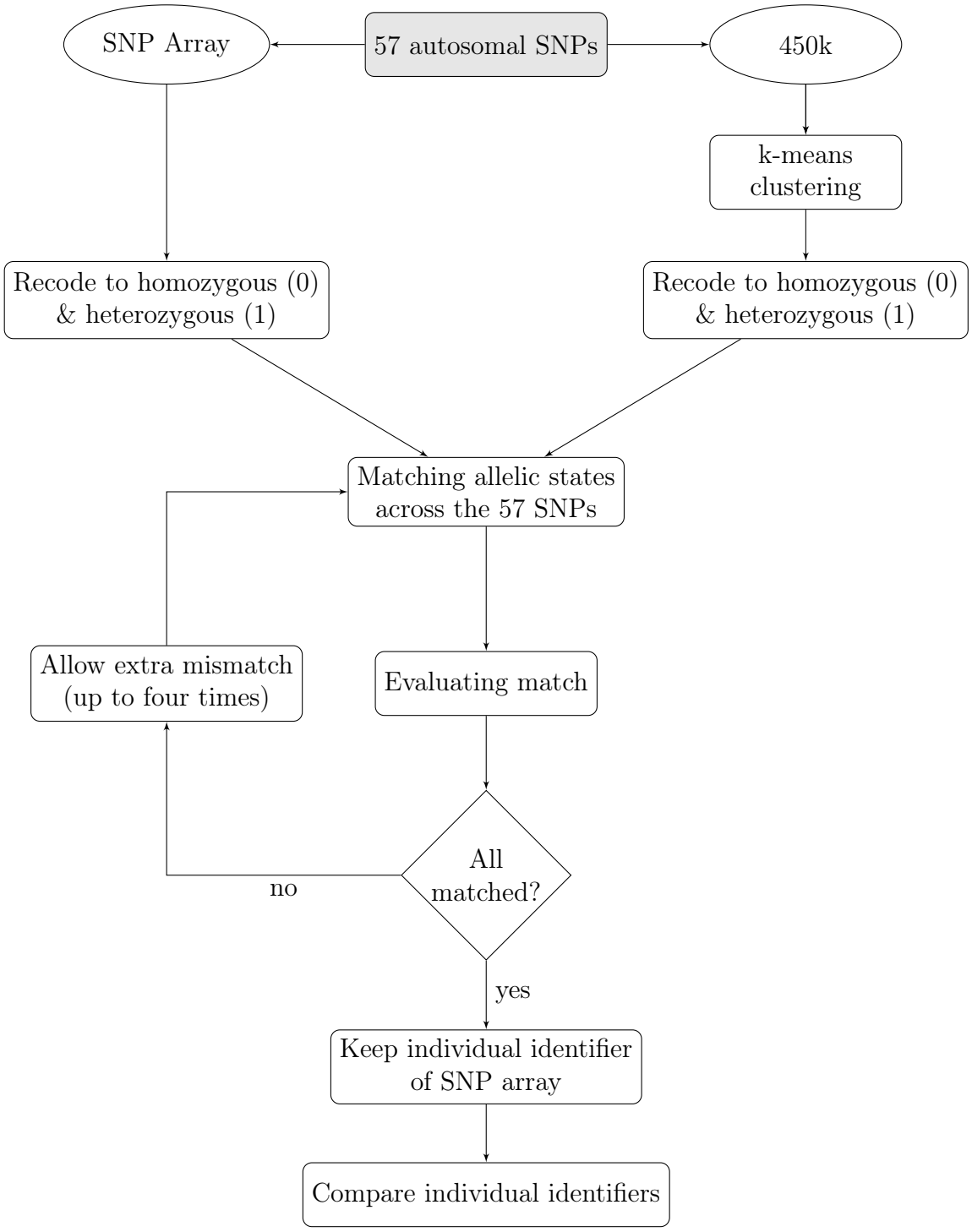



**Figure 2.5: Schematic overview of sample identification.** The 57 SNPs were selected for the known genotype (SNP array) and quality control probes on 450k. These were then recoded to homozygous or heterozygous and assessed for a complete match of all 57 SNP states and selected if 100% match for identification. If not they would be matched again allowing one more mismatch each time and recorded for the least amount of mismatches.

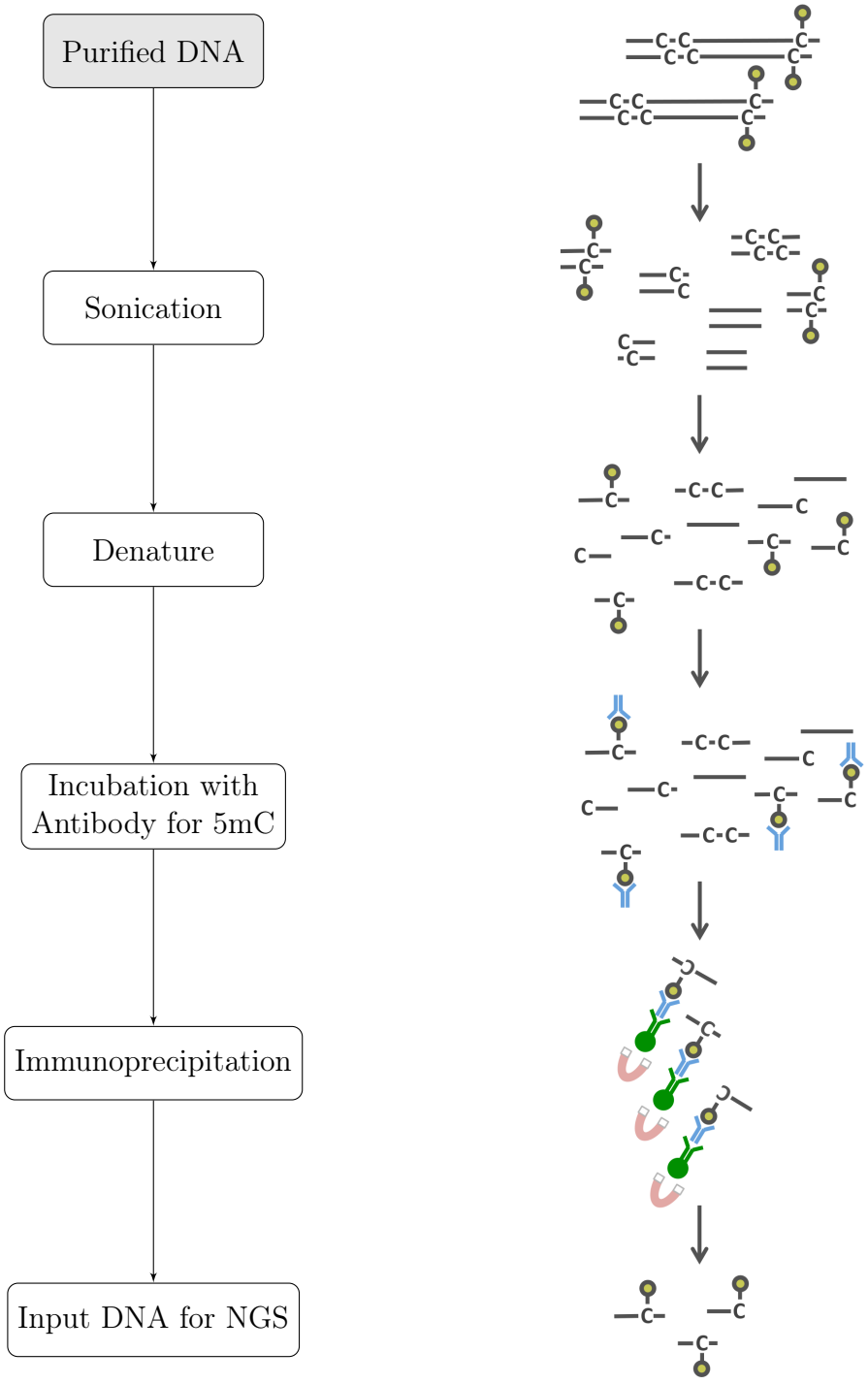



**Figure 2.6: MeDIP-seq design**. Purified DNA is first fragmented by sonication denatured to single strands. Next, the antibodies for 5mC are added followed by immunoprecipitation with antibodies with conjugated magnetic beads. The DNA not bound by 5mC is removed with the supernatant. Finally, the antibodies are digested.

# Chapter 3

# Pan-cancer Biomarkers in Cancer Discordant Monzygotic Twin-pairs

## 3.1 Background

An area of recent interest is the identification of DNA methylation biomarkers that can diagnose cancer in the early stages or identify individuals at risk in non-invasive tissues, such as peripheral blood, serum, or plasma [128, 162, 163]. Whereas serum and plasma can capture free circulating (tumour) DNA, peripheral blood will predominantly capture a DNA mixture from white blood cells. Therefore in a tissue like peripheral blood, the DNA methylome could show systemic changes in the body associated with cancer development or accrued cancer risk factors [37].

Changes in the DNA methylome in peripheral blood have previously been associated with cancerous and pre-cancerous primary locations such as breast [136, 137], colon [138], bladder [139], and ovary [140]. DNA methylation variation has also been associated with major risk factors for cancer such as smoking [68–70, 164, 165], age [33], air pollution [166], and BMI [66]. To date, no study has identified DNA methylome changes in blood samples indicative for various primary locations, or a pan-cancer biomarker. Cancer is a heterogeneous disease but shares characteristics such as uncontrolled cell proliferation and loss of differentiation. Similar changes can been observed across tumours in DNA methylome, the proteome, somatic mutations, and somatic copy number alterations [167–172]. Similarly, there

may be also common systemic changes in individuals at risk or as a response to cancer development. A blood-based DNA methylome biomarker that can predict common cancers or cancer pathogenesis of different locations would be of great benefit for current cancer screening methods. It could also add to our understanding of common systemic changes associated with cancer pathogenesis.

The aim of this chapter was to investigate the DNA methylomes in 41 cancer-discordant female MZ twin-pairs and determine differences in DNA methylation both at single CpG level, pan-cancer associated DMPs (pc-DMPs), as well as at small regional level, pan-cancer associated DMRs (pc-DMRs). DNA methylomes were profiled by the 450k of DNA from blood samples obtained up to five years preceding or up to five years post-cancer diagnosis. The presence of DNA methylation variation prior to diagnosis was assessed as well as biomarker stability by using five additional twin pairs with blood samples obtained up to eleven years preceding diagnosis. The top ranked results were followed up with replication and transcriptomic analyses across three tissues to reveal potential biomarkers.

## 3.2 Methods

### 3.2.1 Sample Selection

Discordant MZ twin-pairs for cancer were selected from one larger dataset comprising 957 DNA methylomes of 915 unique individuals profiled at one centre, the Wellcome Trust Sanger Institute (described in Section 2.3.1). Detailed cancer diagnosis information by UK registered pathologists was accessed through record linkage with the ONS. Discordance was defined as a case where one twin was diagnosed with a (first) single primary site malignant tumour within a five year window of the blood sample extraction, while her co-twin was not diagnosed with any malignant tumour in the most recent records.

This resulted in 41 middle-aged (42 to 79 years of age with a median of 61) female cancer discordant MZ twin-pairs of European descent (see Table 3.1). Individuals were excluded if they were diagnosed with blood and lymph related malignancies, skin cancers apart from melanoma (*i.e.* basal cell carcinoma and squamous cell carcinoma), and premalignant and intraepithelial changes of the cervix. The co-twin not diagnosed with cancer was cancer free in a period ranging from 4 to 21 years subsequent to the diagnosis of the affected co-twin (median = 10.3 years). The 41 MZ twin-pairs included cancers at eight different primary locations: breast (23 pairs), cervix (1 pair), colon (10 pairs), endometrium (1 pair), thyroid gland (1 pair), melanoma (3 pairs), ovary (1 pair), and pancreas (1 pair). The peripheral blood DNA methylome twin-pairs can also be divided into those that were obtained either preceding diagnosis (15 pairs) or post diagnosis (26 pairs).

Major risk factors for cancer pathogenesis were investigated in these 41 MZ twin-pairs: smoking, BMI, and alcohol consumption. Smoking habits were assessed from longitudinal questionnaires and divided into three categories: never smoked, current smokers, and ex-smokers (stopped >3 years before blood sample collection). 29 MZ twin-pairs were concordant in smoking habit: 19 pairs were non-smokers,

1 pair were current smokers, and 9 pairs were ex-smokers. The remaining 12 twin-pairs comprised 7 pairs including an ex-smoker and never smoker co-twin and 5 pairs including an ex-smoker and current smoker co-twin (see Table 3.2). In terms of BMI, the mean BMI of all individuals was 26.9 kg/m$^2$. 21 out of 41 MZ twin-pairs had a greater BMI in the twin diagnosed with cancer compared to her co-twin at time of blood sample. Overall the twin-pairs were very similar in BMI and had a median within-pair difference of 1.7 kg/m$^2$. Three pairs had a difference large than 6 kg/m$^2$, of which the higher BMI was concordant with cancer diagnosis. Finally, self-reported alcohol consumption obtained from longitudinal questionnaires showed no significant discordance within twin-pairs.

**Table 3.1: Characteristics of 41 cancer-discordant MZ twin-pairs.**

| Selection | Characteristic | Mean | Median | Range | |
|---|---|---|---|---|---|
| 41 discordant twin-pairs | Age at DNA extraction | 61.7 | 61.1 | 41.5 | 78.7 |
| | Age at Cancer diagnosis | 60.9 | 60.3 | 43.5 | 75.5 |
| | Cancer Free (yrs) | 10.2 | 10.5 | 5.2 | 22.1 |
| | BMI* | 2.5 | 1.68 | 0.1 | 10.5 |
| 82 Individuals | BMI | 27.3 | 26.3 | 20.4 | 40.6 |

* In absolute differences

**Table 3.2: Smoking habits of 41 cancer-discordant MZ twin-pairs.**

| MZ twin-pair | Never smoked | Ex-smoker | Current smoker |
|---|---|---|---|
| Concordant (number of pairs) | 19 | 9 | 1 |
| Discordant (number of pairs) | 7 | 5 | |

An additional analysis to assess the stability of DNA methylation differences over time used five extra female MZ twin-pairs (see Table 3.3). These pairs had DNA methylomes available 5 to 11 years preceding cancer diagnosis (age range 38-65, median age 57) and were discordant for cancers at primary sites of the breast (3 pairs) and colon (2 pairs). No within twin-pair differences were observed in smok-

ing or alcohol consumption and their BMI ranged from 20.5 – 27.8 kg/m$^2$ (median within-pair difference of 0.6 kg/m$^2$).

**Table 3.3: Characteristics of additional 41 cancer-discordant MZ twin-pairs.**

| Selection | Characteristic | Mean | Median | Range | |
|---|---|---|---|---|---|
| 5 discordant twin-pairs | Age at DNA extraction | 54.3 | 56.6 | 38.1 | 65 |
| | Age at Cancer diagnosis | 62.2 | 63.2 | 43.5 | 75.5 |
| | Cancer Free (yrs) | 7.3 | 8.0 | 2.6 | 11.4 |
| | BMI* | 1.3 | 0.6 | 0.3 | 2.7 |
| 10 Individuals | BMI | 24.1 | 23.3 | 20.5 | 27.8 |

* In absolute differences

### 3.2.2 Genome-wide DNA Methylation Data

Peripheral blood DNA methylomes were profiled with the 450k from bisulphite-converted DNA in two batches of 24 and 68 samples. The array, pre-processing, and QC procedures are described in detail in section 2.3.1.

In short, the summary is described of the quality control procedure and results for the dataset used in this chapter. First probes were removed that: 1) failed detection in one or more samples and/or had a bead count less than 3 in >5% of samples (n = 3,325), 2) aligned to more than one location in the human genome with their 50 bp sequence, 3) located on the sex chromosomes, 4) harboured common genetic variants occurring in European Caucasians (MAF >1%) within 10 bp on the probe at the interrogated CpG site, and 5) contained variants at any MAF at the interrogated CpG site [150, 151]. Therefore the remaining number of probes used for this EWAS was 453,627.

The 41 MZ twin-pairs were verified with the sample identifier (see Section 2.3.1.4) using the 57 autosomal control SNP probes and known genotype data. The beta values were then normalized using the BMIQ method to correct for probe type bias [154].

PCA was performed on standardised beta values (N(0,1)) per probe. The

first 5 PCs combined explained 46.8% of the total variance. No nominally significant association was identified between cancer status and the first five PCs. However, significant associations were identified ($p < 4.1 \times 10^{-3}$) between the first five PCs with all six estimated cell proportions: CD8+ T cells, CD4+ T cells, Natural Killer cells, granulocytes, and monocytes [80], as well as with batch and beadchip.

### 3.2.3 Gene Expression Profiles

Gene expression profiles of LCLs, skin tissue, and adipose tissue from 283 healthy female individuals of European descent of the TwinsUK were obtained from the MuTHER project [146] described in detail in section 2.4. Whole blood DNA methylation in the same 283 healthy female individuals was profiled using 450k, processing and quality control was similar as previously described in section 2.3.1.

### 3.2.4 Statistical Analysis

A general overview of the analyses performed in this chapter is shown in Figure 3.1.

#### Global DNA Methylation Profiles

DNA methylome profiles of the 41 MZ twin-pairs were analysed for global differences using unsupervised hierarchical clustering analysis with Euclidean distances and complete linkage method. This was repeated using only the 1,000 most variable CpG sites across these individuals, identified as the CpG sites with the greatest standard deviations. Within MZ twin-pair genome-wide correlations were performed using a Spearman's rank test. This was compared to pair correlations of individuals randomly assigned to another resulting in 41 new "twin-pairs". The random pairing was repeated 100 times for two subsets; randomly assigned independent of cancer status and randomly assigned but within their cancer status (*i.e.* healthy paired with healthy and cancer affected paired with cancer affected). These groups were

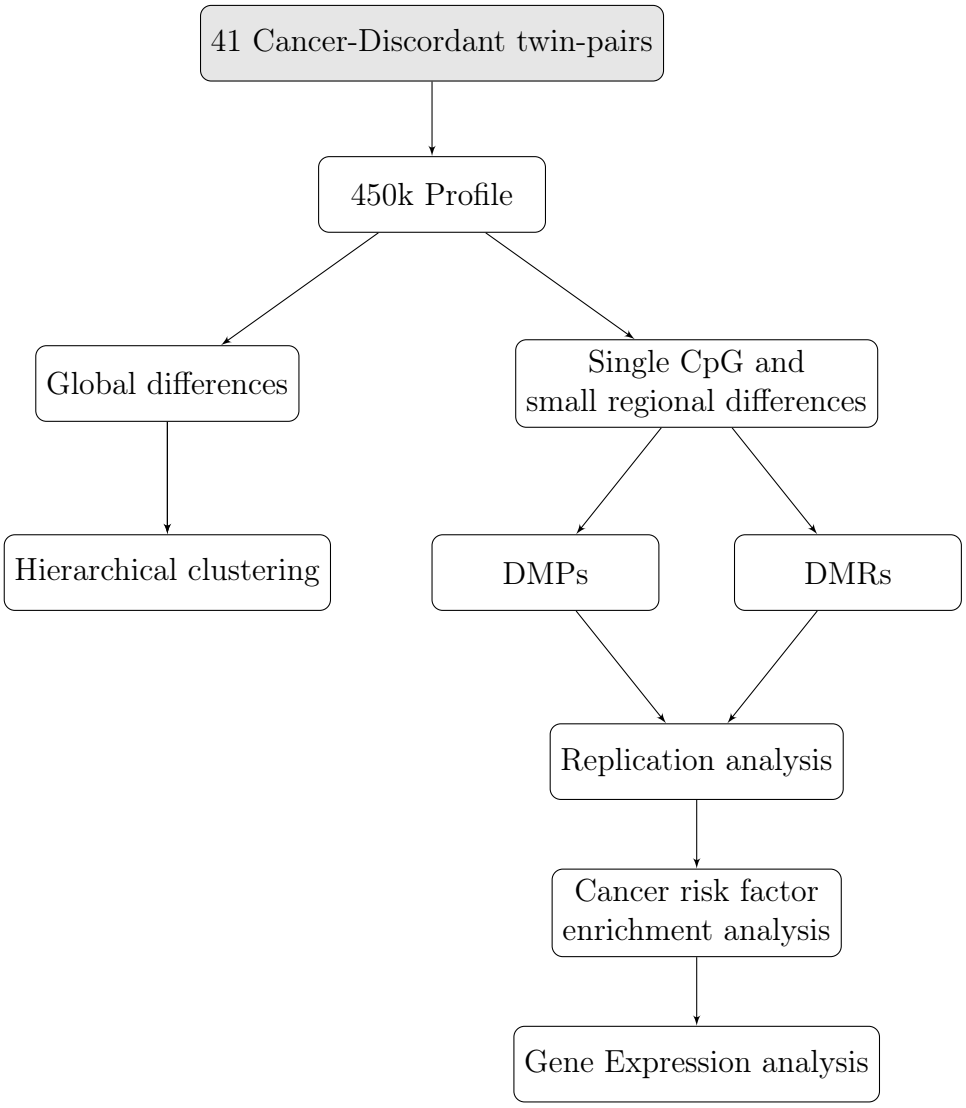



**Figure 3.1: Schematic overview of statistical analyses.** A general overview of the analyses performed in this chapter.

then compared to the true twin-pair group and a two sample t-test was performed for significant differences.

### Pan-cancer Differentially Methylated Positions

The first EWAS was performed at single CpG sites to identify pc-DMPs. Prior to this EWAS, the DNA methylomes were first individually adjusted for confounders. To this end, a linear model was fitted on standardised beta values per probe (N(0,1)) as the response variable and the first five PCs as predictors. These DNA methylation residuals were then used to calculate within twin-pair differences

that were determined consistently as cancer-affected twin minus healthy co-twin. A one-sample t-test was performed on these within-pair differences to assess significance of association. Computational approaches similar to the PC regression applied here to adjust for cell heterogeneity and noise in large-scale epigenomic data sets have already been applied and published in recent years [173–175]. Multiple testing was taken into account by the use of a false discovery rate (FDR) of 10% using the "qvalue" package in R, while suggestive results threshold was set at a nominal $p$ value of 1 x $10^{-5}$.

Another two EWAS were preformed to compare these results to two different pipelines using other methods. First, a linear model was fitted on normalised beta values per probe (N(0,1)) as the response variable and the estimated cell type proportions of granulocytes, CD8+ T cells, and NK cells as well as batch (two levels) as predictors. Again, these DNA methylation residuals were then used to calculate consistently within twin-pair differences (cancer-affected twin minus healthy co-twin) followed by a one-sample t-test to assess significance of association. Second. an EWAS was performed with a computational approach based on surrogate variable analysis (SVA) devised by Houseman *et al.* [173] termed "RefFreeEWAS" using the paired design option: "PairsBootRefFreeEwasModel". The top-ranked probe, still ranked first in all three methods, and the suggestive probes had p-values $<1$ x $10^{-3}$ for all three approaches. The two computational approaches based on PCA and SVA were the most similar as expected.

**Pan-cancer Differentially Methylated Regions**

Next, an EWAS was performed at small genomic regions to identify pc-DMRs. These small regions were predefined and comprised at least 3 CpG sites that were no more than 500 bp apart. To keep the paired structure of the data, differences at single CpG sites within these regions were per MZ twin-pair determined on DNA methylation residuals, similar to the DMP analysis. This was then compared to a

group without DNA methylation differences using the "Bumphunter" package in R [176]. The algorithm determines a $p$ value based here on 1,000 permutations as well as determining family-wise error rate (FWER) adjusted $p$ value. Pc-DMRs were identified as statistically significant at a $p$ value $<0.05$ and at a FWER adjusted $p$ value $<0.5$.

**Cancer Risk Factor Analysis**

Known cancer risk factors, age and smoking, were assessed for enrichment in the top ranked 500 pc-DMPs. To this end, previously published whole blood DMPs associated with age [135] and smoking [69] were selected for enrichment analysis. The enrichment analysis compared the occurrence of these age and smoking DMPs within the top 500 ranked pc-DMPs against their occurrence in the remaining CpG sites on the 450k (n = 453,127). A Fisher's exact test was used to assess significance.

**Gene Expression Analysis**

Gene expression levels of three different tissues were assessed for association with the most associated pc-DMPs and pc-DMR. For 283 healthy individuals, linear mixed effects models were fitted on gene expression levels with age, BMI, batch, concentration (skin tissue only) as fixed effects, and family and zygosity as random effects. A similar linear mixed effects model was fitted on DNA methylation levels in the same individuals with age, BMI, beadchip, position on the beadchip, and granulocytes, monocytes, CD8+ T cells (estimated proportions) as fixed effects, and family and zygosity as random effects. This was followed by a Pearson correlation test on the DNA methylation and expression residuals from these models.

### 3.2.5 Genomic Annotation Analysis

All CpG sites included in this chapter were compared with annotations for CpG density (CGI, shores, and shelves) from the UCSC track [177], RefSeq genes (promoter, 5'UTR end, gene body, 3'UTR, intergenic), and functional genomic

elements derived from ENCODE including ChromHMM state segmentation, DNase-I hypersensitivity sites, and transcription factor binding sites (TFBSs) [178, 179].

For each annotation category an enrichment analysis was performed comparing the top 500 ranked pc-DMP probes to the remainder of CpG sites (n = 453,127). Subsequently, a Fisher's exact test was performed to test significance. For the ChromHMM state segmentation, all 15 states were assessed as well as collapsed states with a single "promoter" and "enhancer" categories. The promoter category comprised active promoter (state 1), weak promoter (state 2), and poised promoter (state 3) and the enhancer category comprised strong enhancer (state 4 and 5), and weak enhancer (state 6 and 7).

### 3.2.6 Replication Sample and Analysis

#### Sample Selection

An independent MZ twin-pair sample from the Netherlands Twin Registry (NTR) with peripheral blood DNA methylomes and white blood cell counts was used for replication of the top results [180]. A detailed official cancer diagnosis in this cohort was similarly obtained through record linkage with the Netherlands Cancer Registry (NKR). After this record linkage, 703 complete MZ pairs who took part in the NTR biobank project remained. Discordant MZ twin-pairs were selected from this set with similar criteria as for the DTR dataset described in section 3.2.1 (see Table 3.4). This resulted in 9 cancer discordant MZ twin-pairs comprised of a mixed sex sample: 4 male and 5 female pairs. These included cancers at 6 different primary locations: breast (3 pairs), meninges (1 pair), pituitary gland (1 pair), prostate (2 pairs), rectum (1 pair), and soft tissue (1 pair).

#### Analysis

Replication of the 4 most associated pan-cancer DMPs was performed in the 9 cancer discordant MZ twin-pairs from the raw 450k output using a similar quality

**Table 3.4: Characteristics of 9 cancer-discordant MZ twin-pairs.**

| Selection | Characteristic | Mean | Median | Range | |
|---|---|---|---|---|---|
| 9 discordant twin-pairs | Age at DNA extraction | 55.7 | 55.5 | 34.6 | 71.9 |
| | Age at Cancer diagnosis | 60.0 | 58.1 | 37.6 | 74.3 |
| | BMI* | 2.7 | 2.8 | 0.3 | 6.0 |
| 18 Individuals | BMI | 26.2 | 26.3 | 19.2 | 33.0 |

* In absolute differences

control, normalisation and analysis pipeline as described for the discovery analysis in this section. In addition, 3 of these pan-cancer DMPs were also available in 480 concordant "healthy" MZ twin-pairs (no cancer diagnosis between them) for which blood samples were extracted at the same date. These three probes were provided by Dr. Jenny van Dongen using her quality control and normalisation pipeline of the complete NTR dataset. The normalisation used in this instance was functional normalisation [155]. These were subsequently investigated for difference in variability in DNA methylation at these sites in the healthy MZ twin-pairs compared to the 9 cancer discordant MZ twin-pairs. For this, absolute within twin-pair differences in normalised beta values were calculated for all pairs and assessed for significance using a Mann-Whitney U test between the groups.

## 3.3 Results

### 3.3.1 DNA Methylomes in Cancer Discordant MZ Twin-pairs

Whole blood DNA methylomes were analysed for 41 female cancer discordant MZ twin-pairs. The individuals were diagnosed with cancer at a single primary site and included cancers at: breast, cervix, colon, endometrium, thyroid gland, skin (melanoma), ovary, and pancreas (see Figure 3.2 A). Overall variation in DNA methylomes was first assessed using unsupervised hierarchical clustering of unadjusted normalised DNA methylation levels. This revealed that the DNA methylomes were not globally different and did not cluster according to cancer status. Thirty-five of the 41 MZ twin-pairs (85.4%) clustered as a pair and the other six MZ twin-pairs clustered according to the beadchip that they were profiled on (see Figure 3.2 B). This highlights the influence of technical confounders for DNA methylome data using the 450k and importance to account for these accordingly in downstream analyses.

To further examine a cancer associated DNA methylation signature, 1,000 most variable CpG sites were selected using the highest standard deviations and were again assessed using unsupervised hierarchical clustering. Now the MZ twin-pairs clustered as twin-pairs and were even more similar with 100% pairing. This indicates that these most variable probes are highly variable between twin-pairs and therefore variation at these probes could likely be due to genetic influences between the different MZ twin-pairs rather than cancer status. Likewise, MZ twin-pairs have stronger within-pair correlations in their DNA methylome than individuals paired at random or paired randomly within affection status category (see Figure 3.2 C, $p = 2.2 \text{ x } 10^{-16}$) representing the strong influence of genetics on these profiles. The average correlation within MZ twin-pairs (Spearman's rank correlation coefficient ($r_S$) = 0.986) is comparable to previous estimated genome-wide correlations in new-born twins, ranging from 0.98-0.99 (placenta, cord

blood mononuclear cells, and human umbilical vascular endothelial cells) [181], 0.99 at 15 years (peripheral blood) [101], and 0.98 and 0.99 in middle-aged individuals (peripheral blood, adipose tissue respectively) [143, 182].

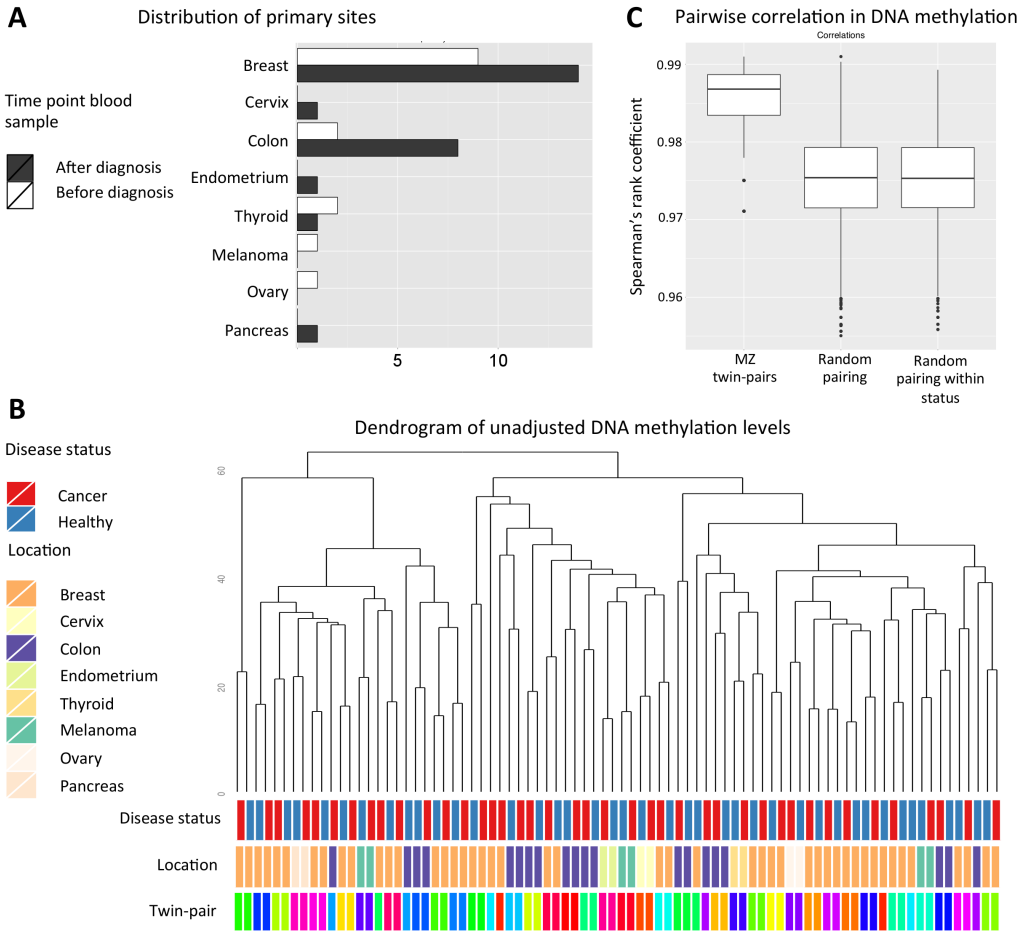



**Figure 3.2: Diagnostic characteristics and global DNA methylation profiles of 41 cancer-discordant MZ twin-pairs.** **(A)** Number of cases for each individuals primary cancer location with blood samples obtained preceding (white) or after (black) cancer diagnosis. **(B)** Dendrogram of unadjusted DNA methylomes depicting annotation bars for affection status per individual, and primary cancer location as well as family identifier that coloured per twin-pair. **(C)** Pair-wise correlation of DNA methylomes show greater similarity within MZ twin-pairs compared to pairs of unrelated individuals, either paired at random or within disease status.

### 3.3.2 Pan-cancer Associated Differentially Methylated Positions (DMPs)

Next, differential DNA methylation levels at single CpG sites were analysed epigenome-wide for pc-DMPs within the 41 MZ twin-pairs discordant for cancer diagnosis. The DNA methylomes were adjusted prior to the analysis for the first five PCs. These explained 46% of the variance across the 41 MZ twin-pairs and were not associated with affection status ($p$ >0.5), but were significantly associated with beadchip, batch, and estimated blood cell type proportions of all six cell types (see Section 3.2.2). The EWAS was then performed using a one-sample t-test on the directional within-pair differences of the adjusted DNA methylation data.

One novel pc-DMP was identified at a FDR threshold of 10% for the intergenic probe cg02444695 ($p$ = 1.8 x $10^{-7}$, see Table 3.5 on page 74, and Figure 3.3 A). The adjusted DNA methylation levels at this pc-DMP were consistently higher in the cancer-affected twins compared to the healthy co-twins. This directional difference was also observed in the normalised unadjusted DNA methylation levels at an average of 0.7% within twin-pairs with a range of -0.9% to 3.0% (see Figure 3.3 B). The CpG is ~70 kb upstream of *SASH1*, which is the nearest gene. *SASH1* is a tumour suppressor that is linked to metastasis formation in different types of cancer [183–185].

A further three suggestive associations were observed ($p$ <1.0 x $10^{-5}$) for probes cg26079695 in *COL11A2*, cg27094856 in *AXL*, and cg21046959 in *LINC00340* (see Figure 3.3 A and Table 3.5 on page 74). The CpG interrogated by cg27094856 is in the fourth intron of *AXL* and the expression of this gene has been associated with cancers of various primary sites and stages as well as being a therapeutic target for antibody based therapies [186–188]. The CpG site at cg21046959 is located within the long non-coding RNA *LINC00340*, which has been associated via GWAS and epigenetically with both neuroblastoma and ovarian tumours respectively [189, 190].

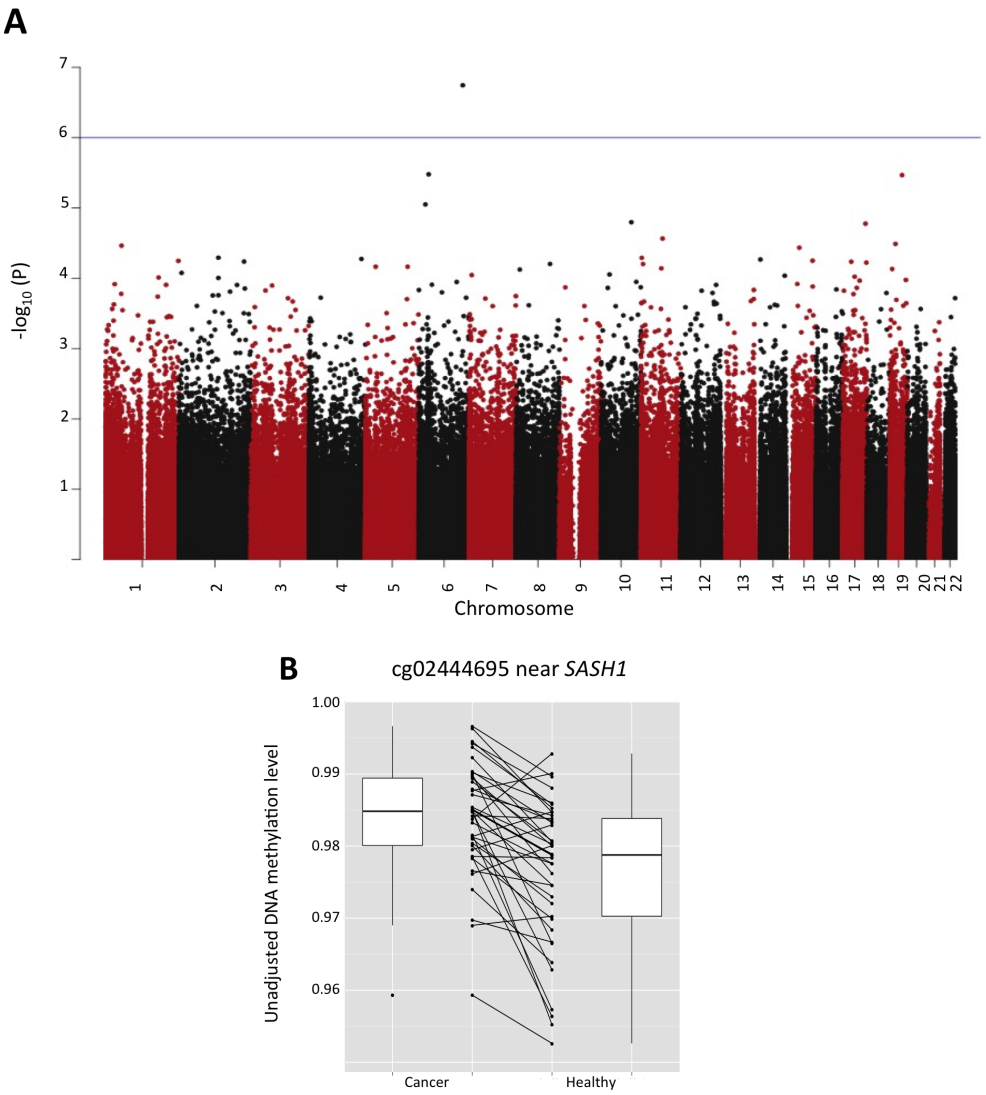



**Figure 3.3: Pan-cancer EWAS results in 41 discordant MZ twin-pairs. (A)** Manhattan plot of EWAS results in which each point depicts the observed -$\log_{10}$ $p$ value per CpG site. **(B)** Association between MZ twin-pairs at the top-ranked CpG cg02444695 near *SASH1*. Normalised unadjusted beta values are shown of cancer-affected individuals (left) and healthy individuals (right) with the lines connecting each MZ twin-pair.

#### 3.3.2.1 Top Pan-cancer DMPs in an Independent Sample

The four most associated pc-DMPs were assessed in an independent MZ twin-pair sample from the NTR. A total of nine middle-aged cancer discordant MZ twin-pairs were selected with the same criteria as the discovery samples, though the NTR twins comprised both male (4) and female (5) twin-pairs. The MZ twin-pairs were first analysed for replication in a similar fashion to the discovery EWAS. The direction of association at three CpG sites was comparable, with the exception of cg02444695 (see Table 3.5). However, no nominal significance ($p$ <0.05) was reached which could be due to the small size of the sample.

Furthermore, an extra 480 healthy MZ twin-pairs of the NTR were also assessed for variability in DNA methylation at these most associated CpGs compared to the nine cancer-discordant pairs. The hypothesis behind these analyses was that if the observed pc-DMPs were associated with affection status, less variation at these sites would be expected in a healthy population. Three of the four top-ranked probes (g02444695, cg26079695, and cg27094856) were available in the NTR cohort, and the absolute within-pair differences in DNA methylation were compared for the healthy and cancer-discordant groups of MZ twin-pairs (see Table 3.5 and Figure 3.4). Greater variation was observed in cancer-discordant MZ twin-pairs compared to the healthy twin-pairs. At cg27094856 in *AXL* this was a significant difference ($p$ = 0.047), with a healthy median of 0.78% vs cancer median of 1.44% DNA methylation within-pair difference. Additionally, a trend of potential difference was observed at cg02444695 near *SASH1* with a healthy median 1.48% vs. cancer median 2.32% DNA methylation difference ($p$ = 0.091).

**Table 3.5: Top-ranked results of pan-cancer EWAS of 41 discordant MZ twin-pairs.**

| | | | | **Discovery EWAS** n = 41 | | **Replication** n = 9 NTR | | **Variability** n = 9 vs n = 480 NTR | **Prior to diagnosis EWAS** n = 15 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CpG | Position (hg19) | Gene | Location | Mean difference* | *p* value | Mean difference* | *p* value | *p* value | Rank | Mean difference* | *p* value |
| cg02444695 | Chr6:148950185 | - | - | 0.70 | $1.8 \times 10^{-7}$ | -0.64 | 0.26 | 0.09 | 10 | 0.88 | $2.4 \times 10^{-5}$ |
| cg26079695 | Chr6:33143273 | *COL11A2* | Intron | -0.67 | $3.3 \times 10^{-6}$ | -050 | 0.23 | 0.34 | 1518 | -0.88 | $4.1 \times 10^{-3}$ |
| cg27094856 | Chr19:41732589 | *AXL* | Intron | 0.56 | $3.4 \times 10^{-6}$ | 0.02 | 0.96 | 0.05 | 3801 | 0.51 | $9.7 \times 10^{-3}$ |
| cg21046959 | Chr6:22180833 | *LINC00340* | Transcript | -0.53 | $8.9 \times 10^{-6}$ | -0.43 | 0.37 | - | 407 | -0.73 | $1.2 \times 10^{-3}$ |

*The mean differences are calculated as cancer - unaffected co-twin using adjusted DNA methylation values.

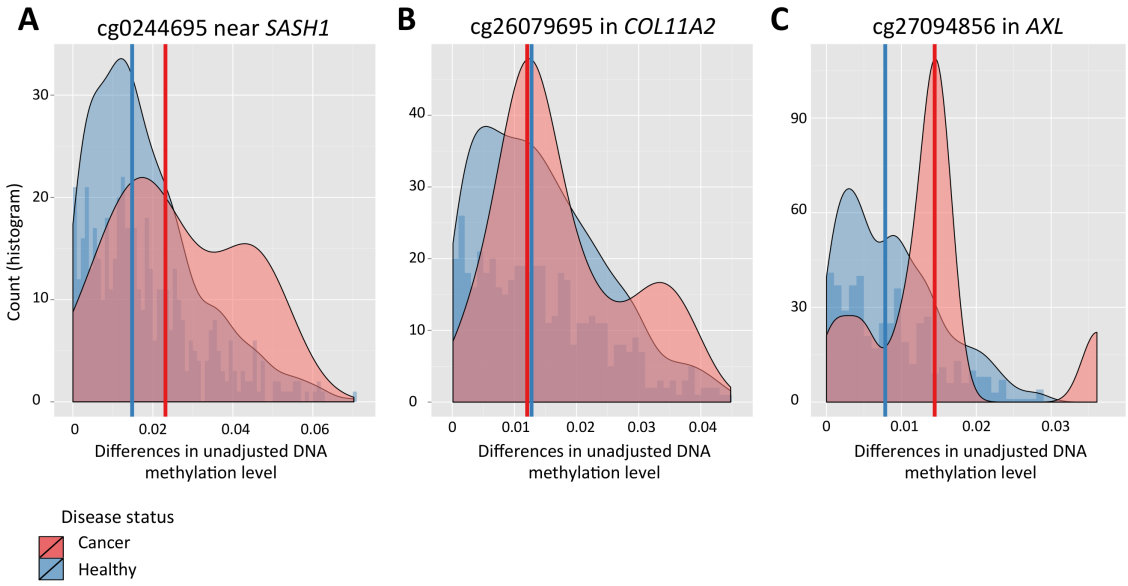



**Figure 3.4: Variability at three top-ranked pan-cancer DMPs in 9 cancer discordant and 480 healthy MZ twin-pairs from the NTR.** Histograms with density overlay with median of absolute differences for 480 healthy MZ twin-pairs (blue) and density with median of 9 cancer discordant MZ twin-pairs (red). At **(A)** cg0244695 near *SASH1*, **(B)** cg26079695 in *COL11A2*, and **(C)** cg27094856 in *AXL*.

### 3.3.3 Pan-cancer Differentially Methylated Regions (DMRs)

Next, small genomic regions encompassing multiple CpG sites associated with affection status, pc-DMRs, were investigated in the sample of 41 MZ twin-pairs. To this end, regions were defined to contain at least three CpGs no more than 500 bp apart. At each region the DNA methylation difference at each CpG was determined per MZ twin-pair (cancer affected twin minus healthy co-twin) using PC-adjusted DNA methylation. Then, the peak-calling algorithm 'Bumphunter' [176] was used to identify pc-DMRs.

This analysis identified one suggestive pc-DMR at the TSS of *TIMM44* that spans ~1 kb (chr19:8,008,080-8,009,137 (hg19), $p = 0.01$, Figure 3.5). The peak of the region is formed by a single CpG, cg14044916 at chr19:8,008,850, and its two neighbouring CpGs that exhibit higher DNA methylation in the cancer-affected

twins. Cg14044916 was ranked 24$^{th}$ in the single CpG EWAS ($p$ = 7.38 x 10$^{-5}$). The pc-DMR overlaps a 5' CGI and this region shows active promoter evidence from ChromHMM in various tissue cell lines of ENCODE [178, 179]. *TIMM44* itself has previously been associated with familial non-medullary thyroid carcinoma [191], aggressive serous ovarian cancers [192] and breast cancer recurrence [193].

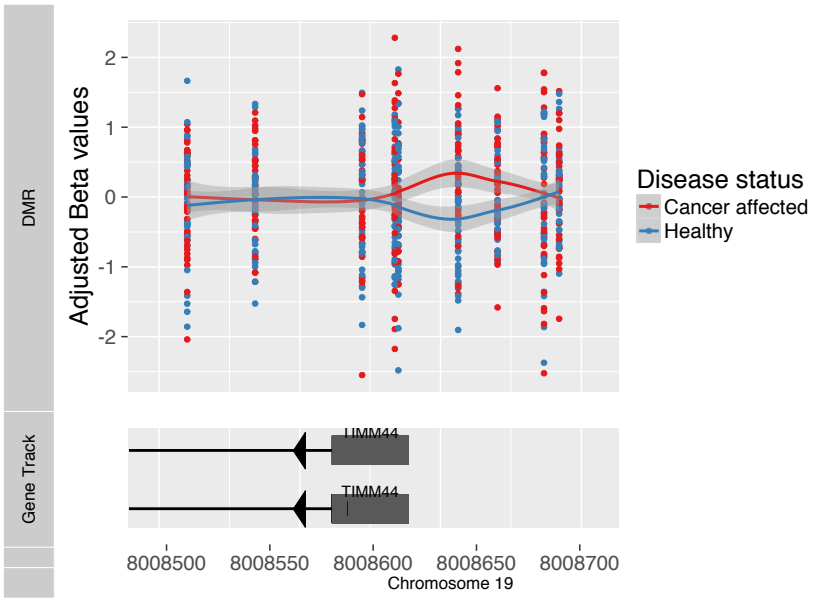



**Figure 3.5: Pan-cancer DMR at *TIMM44*.** Adjusted DNA methylation values at each CpG site in the DMR and smooth (LOESS) lines are shown for individuals affected by cancer (red) and healthy co-twins (blue). In the lower part of the figure is the genomic position (hg19) of the RefSeq genetrack.

### 3.3.4 No Enrichment for Cancer Risk Factors Smoking and Age

The 500 most associated CpGs were assessed for enrichment for two major risk factors for cancer: age and smoking. For this purpose, published peripheral blood DMPs for age and smoking were obtained and examined for co-occurrence of

the pc-DMPs. Age enrichment analysis was performed using the age DMPs from Steegenga *et al.* [135] and their combined table of age DMPs from eight prior studies, 7,318 of which were present in this data. Within the top 500 most associated pc-DMPs, no enrichment was observed and only eight pc-DMPs were also associated with age ($p = 1$). Smoking enrichment was assessed with published smoking DMPs from the largest study to date with the 450k by Zeilinger *et al.* [69]. Out of these, 948 CpG sites were present in our data. Again no enrichment was observed in the top 500 pc-DMPs (three CpGs, $p = 0.089$). Finally, none of the four most associated pc-DMPs have been previously associated with smoking status or age.

The DNA methylome data used in these analyses was adjusted by using the first five PCs. The PCs were not significantly associated with either age or smoking alone, but they may still account for some of the variation attributed to these factors. Subsequently, an EWAS was also performed on DNA methylome data that was only adjusted for batch effects and cell counts (see Section 3.2.4). These top 500 results were similarly assessed and showed no enrichment for age and smoking DMPs. Collectively, the pc-DMPs seem to indicate a more complex portrait of these risk factors or disease biology in the peripheral blood DNA methylome.

### 3.3.5 Biomarker Potential: Analysis in Samples Obtained Preceding Diagnosis

To assess biomarker potential for early diagnosis in this dataset, a subset was selected of 15 MZ twin-pairs where the DNA methylome blood samples were obtained up to five years preceding diagnosis (see Figure 3.2 A). A similar EWAS adjusting for the first 5 PCs in these 15 MZ twin-pairs revealed additional pc-DMPs including at the promoters of *COX7C* and *U2AF1* in the 10 top ranked results (see Table 3.6 on page 80). In particular, the CpG in *COX7C* (cg04533633, $p = 3.0 \times 10^{-6}$) is in the same region earlier identified by Marsit *et al.* [194], using the 27k, as one of the nine most significant loci in peripheral blood associated with

cancer of the bladder. At this region, a similar direction of effect is seen with higher DNA methylation in the cancer-effected individual. In *U2AF1*, recurrent somatic mutations have been identified across tumour tissue as pan-cancer mutations. These mutations were shown to induce changes in the transcriptome through differential splicing [195].

The top-ranked pc-DMP identified in the main analysis of 41 MZ twin-pairs, cg02444695, was now ranked tenth in this EWAS and thus still highly significant with same direction of effect ($p$ = 2.40 x $10^{-5}$, see Table 3.6 and Figure 3.6 A). The three suggestive pc-DMPs of the main EWAS remained significant in the subset EWAS of samples prior to diagnosis (see Table 3.5). In fact, for cg02444695, cg26079695, and cg21046959 a greater DNA methylation residual difference was observed within these MZ twin-pairs with blood samples obtained prior to diagnosis, compared to the differences across all 41 MZ twin-pairs. The reduced significance is likely due to the lower number of MZ twin-pairs included in this EWAS.

### 3.3.6 Pan-cancer Biomarker Stability Over Time

The four most associated pc-DMPs as well as the pc-DMP at *COX7C* were investigated in depth to assess the relationship between blood sample collection in respect to diagnosis and DNA methylation differences. For the top-ranked cg02444695 (near *SASH1*), the largest difference in DNA methylation is seen when the blood was extracted around the year of diagnosis. Whilst for cg26079695 in *COL11A2* and cg21046959 in *LINC00340*, ranked second and fourth respectively, the greatest differences were observed in samples obtained within the 5 year period before diagnosis. The third ranked probe was the only one with greatest difference after diagnosis (see Figure 3.6 B). The age of diagnosis and age at blood sample extraction were not significantly correlated ($p$ = 0.29, see Figure 3.6 C). Thus increased variation with chronological age seem less likely to explain these pc-DMPs. This is also illustrated in Figure 3.6 D for the most associated pc-DMP, that shows

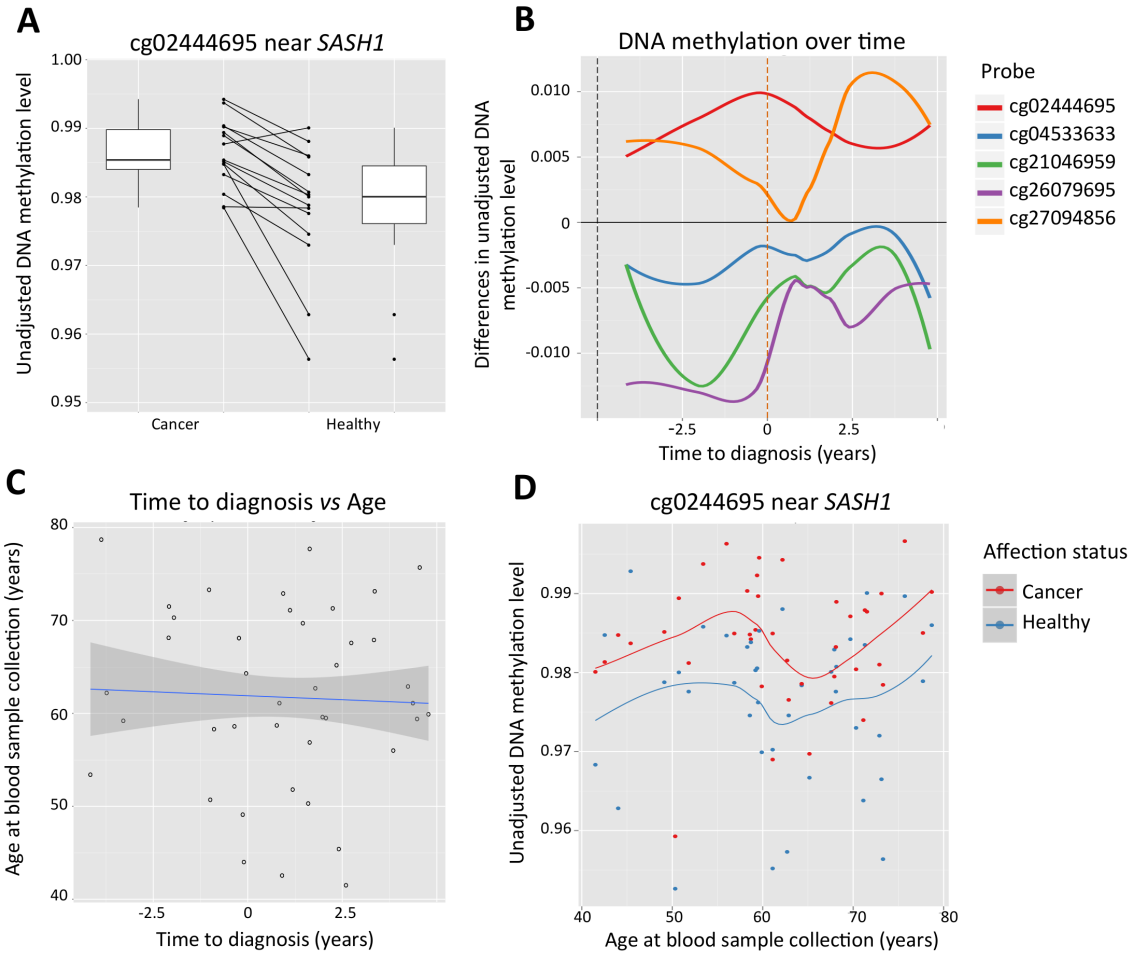



**Figure 3.6: Differential DNA methylation at time of cancer diagnosis and age.** **(A)** Association between 15 MZ twin-pairs at the top-ranked CpG cg02444695 near *SASH1* for samples obtained prior to diagnosis. Normalised unadjusted beta values are shown of cancer-affected individuals (left) and healthy individuals (right) with the lines connecting each MZ twin-pair. **(B)** Directional differences in adjusted DNA methylation values within twin pairs at the four top-ranked pc-DMPs and cg04533633 (at *COX7C*) represented by smooth (LOESS) lines (see legend). **(C)** Years to diagnosis compared to age at blood sample collection with a least squares regression fit line. **(D)** Unadjusted normalised DNA methylation values at cg02444695 (near *SASH1*) in affected individuals (red) and healthy co-twins (blue) at age of blood sample extraction with smooth (LOESS) lines.

**Table 3.6: Top-ranked results from the EWAS of 15 MZ twin-pairs prior to diagnosis.**

| | | | | **EWAS prior to diagnosis** n = 15 | | | **Discovery EWAS** n = 41 | | |
|---|---|---|---|---|---|---|---|---|---|
| CpG | Position (hg19) | Gene | Location | Rank EWAS | Mean difference* | *p* value | Rank EWAS | Mean difference* | *p* value |
| cg22786903 | chr14:24912113 | *SDR39U1* | TSS 200 | 1 | -0.86 | $2.8 \times 10^{-6}$ | 50680 | -0.26 | 0.11 |
| cg04533633 | chr5:85913706 | *COX7C* | TSS 200 | 2 | -0.75 | $3.0 \times 10^{-6}$ | 279 | -0.5 | $7.2 \times 10^{-4}$ |
| cg02272547 | chr16:89634031 | - | - | 3 | 0.82 | $5.0 \times 10^{-6}$ | 390 | 0.45 | $1.0 \times 10^{-3}$ |
| cg07884474 | chr20:2821816 | *VPS16*;*FAM113A* | Body;TSS 1500 | 4 | -0.79 | $6.0 \times 10^{-6}$ | 518 | -0.36 | $1.4 \times 10^{-3}$ |
| cg04479472 | chr7:1231443 | - | - | 5 | 0.56 | $1.6 \times 10^{-5}$ | 20627 | 0.21 | 0.05 |
| cg27581871 | chr16:542535 | *RAB11FIP3* | Body | 6 | 0.73 | $1.8 \times 10^{-5}$ | 185567 | 0.13 | 0.41 |
| cg14296488 | chr21:44527812 | *U2AF1* | TSS 200 | 7 | 0.69 | $2.0 \times 10^{-5}$ | 1331 | 0.43 | $3.4 \times 10^{-3}$ |
| cg10305789 | chr20:37434167 | *PPP1R16B* | TSS 200 | 8 | -0.5 | $2.1 \times 10^{-5}$ | 43093 | -0.19 | 0.1 |
| cg07178008 | chr18:19445471 | *MIB1* | 3'UTR | 9 | -1.35 | $2.3 \times 10^{-5}$ | 130784 | -0.22 | 0.29 |
| cg02444695 | chr6:148950185 | - | - | 10 | 0.88 | $2.4 \times 10^{-5}$ | 1 | 0.7 | $1.8 \times 10^{-7}$ |

*The mean differences are calculated as cancer - unaffected co-twin using adjusted DNA methylation values.

stable differences across all ages.

Further investigation of early DNA methylation differences within MZ twin-pairs included an extra five cancer discordant MZ twin-pairs with blood samples between 5 and 11 years preceding diagnosis. This revealed that there was not a clear signal during these earlier years for the top associated CpGs. At cg02444695 particularly, a small reversed pattern is seen between 5 and 11 years compared to 5 year range preceding diagnosis. Concluding that the DNA methylation differences can be observed up to five years before the official cancer diagnosis.

### 3.3.7 Functional Follow Up of Pan-cancer Differential Methylation Results

Association with gene expression of the nearest genes was assessed for the four most associated pc-DMPs and pc-DMR. To this end, 283 healthy female individuals were selected with peripheral blood DNA methylomes and transcriptomic profiles from three tissues: LCLs, skin, and adipose. The correlation was assessed between DNA methylation and the nearest available expression levels and identified two significant correlations ($p$ <0.05, see Table 3.7). The first correlation was observed at cg21046959 with the closest protein-coding transcript of *PRL* in LCLs, ~100 kb upstream of the CpG (r = 0.17, $p$ = 4.5 x $10^{-3}$, see Figure 3.7 A). The second correlation was observed at cg27094856 with expression of *AXL* in skin tissue (r = -0.15, $p$ = 0.01, see Figure 3.7 B). Both these two correlations were not observed across all tissues. The CpG positively correlated to *PRL* expression was annotated by ENCODE in heterochromatin in the GM12878 B-lymphocyte cell line. However, it was annotated as an active promoter and weak enhancer in human embryonic stem cell line (H1-hESC) and leukaemia cell line (K-562), respectively. The CpG negatively correlated to the expression of *AXL* in skin tissue was annotated as a strong enhancer in epidermal keratinocytes (NHEK) and as an inactive or poised promoter in the GM12878 cell line.

Next, the top 500 CpGs most associated with cancer affection status were assessed for enrichment of functional annotations to test for systemic changes that occur in individuals due to cancer pathogenesis. The top 500 CpGs were compared to the remaining CpGs in the dataset (n = 453,127) and an enrichment was identified for pooled enhancers from ENCODE (GM12878, $p = 0.030$, see Figure 3.7 C). State 7 "Weak enhancer" category by ChromHMM ($p = 0.034$) was the only one out of the 4 states pooled together that was nominally significant itself in the enrichment analysis. There was a trend towards depletion of DNA methylation in CGI shores, repressed regions, and weakly transcribed regions, but these were not nominally significant.

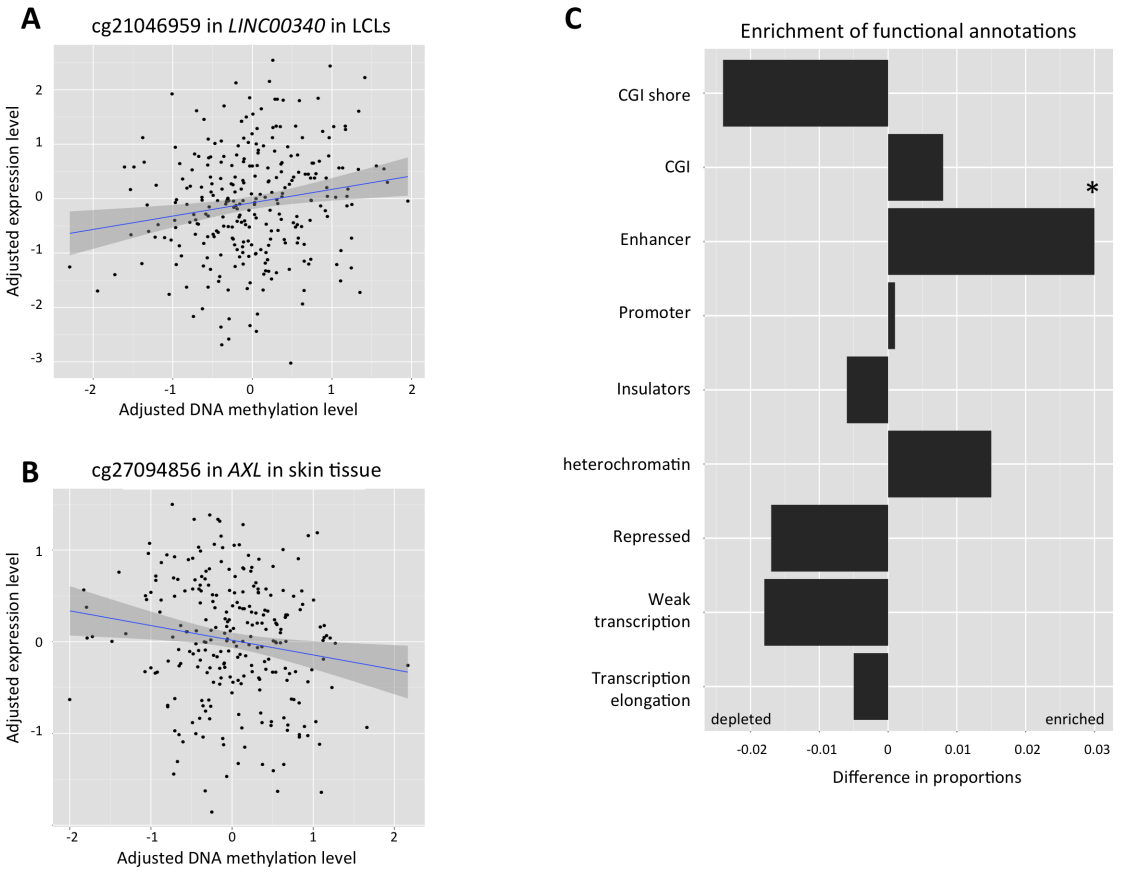



**Figure 3.7: Functional follow up of top-ranked pan-cancer DMPs.** Adjusted DNA methylation levels of peripheral blood compared to adjusted gene expression levels showing a least squares regression fit for 283 individuals at **(A)** cg21046959 and ILMN_1809352 (*PRL*) in LCLs, and **(B)** cg27094856 and ILMN_1701877 (*AXL*) in skin tissue. **(C)** Enrichment of genomic annotation categories within the 500 top-ranked pc-DMPs. The difference in proportion of pc-DMPs compared to the remainder of probes in the genomic annotation classes is depicted by the bars. Nominally significant results were obtained for the "enhancer" category ($p = 0.03$).

**Table 3.7: Gene expression analysis of top ranked pan-cancer DMPs and DMR in 283 individuals.**

| | | | | | | | **LCLs** | | **Adipose tissue** | | **Skin tissue** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CpG | Position (hg19) | Gene | Location | Nearest transcript probe | Gene | CpG distance | r | *p* value | r | *p* value | r | *p* value |
| cg02444695 | Chr6:148950185 | - | - | ILMN_2185984 | SASH1 | 77 kb upstream | -0.02 | 0.79 | 0.01 | 0.91 | 0.03 | 0.59 |
| cg26079695 | Chr6:33143273 | *COL11A2* | Intron | ILMN_2311456 | *COL11A2* | - | 0.04 | 0.46 | 0.07 | 0.21 | 0.09 | 0.17 |
| cg27094856 | Chr19:41732589 | *AXL* | Intron | ILMN_1701877 | *AXL* | - | 0.01 | 0.91 | 0.07 | 0.20 | -0.15 | 0.01 |
| cg21046959 | Chr6:22180833 | *LINC00340* | Transcript | ILMN_1809352 | *PRL* | 106 kb downstream | 0.17 | $4.5 \times 10^{-3}$ | 0.03 | 0.59 | 0.02 | 0.74 |
| cg14044916 | Chr19:8,008,850 | *TIMM44* | TSS 1500 | ILMN_1784031 | *TIMM44* | - | 0.02 | 0.75 | 0.05 | 0.44 | 0.04 | 0.55 |

## 3.4 Discussion

This chapter investigated 41 cancer discordant MZ twin-pairs to identify a pan-cancer differential DNA methylation change in peripheral blood independent of host genetic variation. One epigenome-wide significant pc-DMP (FDR 10%) was identified ~70 kb upstream of *SASH1*. A further three pc-DMPs were identified that passed a suggestive significance threshold that were located within the genes *COL11A2*, *AXL*, and in *LINC00340*. Three out of these four pc-DMPs showed the greatest differences in twin-pairs sampled within 5 years prior to and around time of diagnosis (pc-DMPs near *SASH1*, in *COL11A2* and *LINC00340*). An additional pc-DMP in the promoter of *COX7C* was identified in an EWAS using subset of DNA samples obtained only preceding the cancer diagnosis. This pc-DMP was in the same locus previously associated in peripheral blood samples with bladder cancer [194]. Through a regional-based approach, one pc-DMR at the TSS of *TIMM44* was also identified in the 41 MZ twin-pairs. No overall difference in DNA methylomes was observed for affection status. Finally, MZ twin-pairs exhibited greater within-pair correlation than when individuals paired at random, with correlation levels similar to previous estimates in healthy MZ twin-pairs [93, 101, 143, 181, 182].

The most associated pc-DMP (cg02444695 near *SASH1*) had consistently greater DNA methylation levels in the twin with cancer compared to the healthy co-twin. Using unadjusted DNA methylation betas, this was quantified as a mean of 0.7% within a range -0.9% to 3.0% DNA methylation. The CpG site is in a weakly transcribed region and ~500 bp upstream of a weak/poised enhancer identified by ChromHMM in GM12878 B-lymphocyte cell line (LCLs). The nearest downstream gene is *SASH1* and its expression is associated in various different tumour tissues, such as breast, colon, and bone, with increased metastasising ability and aggressive tumour growth [183–185]. Indeed, a weak negative correlation at cg02444695 was identified between peripheral blood DNA methylation at cg02444695 and ex-

pression levels of *SASH1* in LCLs in 283 healthy individuals. It should be noted that the transformation of lymphocytes to an unusually long life span by EBV can cause less transcriptomic variability [196] and impact the DNA methylome in the mature LCLs [197]. Nevertheless, consistent transcriptomic changes in LCLs have been associated with smoking for example [198]. Near significant increased variability of DNA methylation at this CpG was observed for an independent sample of nine cancer-discordant MZ twin-pairs compared to 480 healthy MZ twin-pairs not diagnosed with cancer to date. The pc-DMP did not directly replicate in these nine twin-pairs, which may in part be due to the small sample size and/or gender composition of four male and five female pairs. The greatest within-pair difference of DNA methylation was observed around the time of diagnosis. This differential effect was detected specifically in samples preceding cancer diagnosis, early or prior cancer pathogenesis, and could reflect accrued risk factor exposures, systemic response, or even surrogate changes. The observed difference was observed to be limited to approximately five years prior to diagnosis.

One of the suggestive pc-DMPs, (cg27094856 in *AXL*), showed consistent increased levels of DNA methylation in the twins with cancer compared to their healthy co-twins. This was also observed in the independent sample of nine cancer-discordant MZ twin-pairs although this was not nominally significant. There was, however, a significant greater variability between the nine cancer-discordant MZ twin-pairs compared to 480 healthy MZ twin-pairs not diagnosed with any type of cancer. This pc-DMP is in the fourth intron of *AXL* and a region annotated by ChromHMM as inactive or poised promoter in GM12878. *AXL* expression is associated with therapy resistance, proliferation, and migration capacity as well as a therapeutic target [186–188]. Transcriptomic analyses identified a negative correlation between DNA methylation at this site with expression levels of *AXL* in skin tissue only. The greatest differences in DNA methylation are observed in samples obtained post diagnosis in which effects caused by treatment cannot be

ruled out.

DNA methylation levels of another suggestive pc-DMPs, cg21046959 in *LINC00340*, were decreased in the twins with cancer compared to their healthy co-twins. A similar trend was found in the replication sample of nine discordant MZ twin-pairs. This site was not available for the variability analyses performed on the other CpG due to quality control for the entire sample performed by the NTR. The CpG is positioned within the last intron of *LINC00340* in a heterochromatin block in GM12878, however this region is an active promoter in K-562, a leukaemia cell line. *LINC00340* itself has been implicated as a susceptibility locus for neuroblastoma as well as showing hypermethylation at its promoter in clear cell ovarian tumours [189, 190]. Gene body hypomethylation coupled with hypermethylation at a promoter are classically associated with lower expression, which fits with these findings [199]. Unfortunately, expression levels for were not available *LINC00340*, although the nearest transcript of *PRL* showed a significant positive correlation in LCLs with DNA methylation at this site. Conversely, *PRL* expression has been positively associated with tumour progression across different cancer types [200–202]. Again, the greatest differences were observed in the samples preceding diagnosis within 5 years, with minimal differences in the 5 years post diagnosis.

The second most associated and suggestive pc-DMP, cg2607969 in *COL11A2*, had decreased DNA methylation in the cancer-affected twin compared to the healthy co-twin. This direction of effect was also observed in replication MZ twin-pairs, although, the CpG site did not show increased variability in the cancer discordant pairs. Located in the gene body of *COL11A2*, *COL11A2* has not been linked to cancer yet. The strongest differences are observed within five years prior to diagnosis.

The EWAS in a subset of 15 MZ twin-pairs with blood samples obtained prior to diagnosis further identified differential DNA methylation at a CpG,

cg04533633, in the promoter of *COX7C*. The same locus in peripheral blood was earlier already associated with bladder cancer [194]. None of the MZ twin-pairs here however were diagnosed with bladder cancer to date. This requires more follow up to investigate if the pan-cancer effects here would extend to bladder tumours as well.

The regional analysis identified a pc-DMR in the promoter of *TIMM44* that exhibits higher DNA methylation levels in the cancer-affected twins. It overlaps a 5' CGI and is annotated by ChromHMM as an active promoter in all of the primary tissue cell lines of ENCODE. *TIMM44* germline genetic variants have been linked to familiar non-medullary thyroid carcinoma [191] and decreased DNA methylation in one of its two intragenic CGIs has been found in aggressive serous ovarian cancers [192]. Additionally, increased expression of *TIMM44* is associated in breast cancers with recurrence after chemotherapy [193]. This region is a good candidate for further assessment with with higher regional resolution technologies, such as targeted bisulphite sequencing.

One of the limitations of this study is the heterogeneous nature of cancer spanning many tissues with varying aetiology. This heterogeneity has the potential to mitigate distinct cancer type effects and thereby reducing the power to detect true differential DNA methylation associations. On the other hand, pan-cancer signatures in various biogoical processes as well as certain "driver" mutations observed across cancer tissues [168–172] show that potential general systemic effects or even surrogate effects could occur within individuals.

Another potential limitation is that even though blood is relatively non-invasive and easy to obtain tissue, it is a very heterogeneous tissue. This cellular heterogeneity is here captured by using PCs that are correlated with the major estimated cell proportions. Nevertheless, this or any statistical approach cannot fully correct the data for cell heterogeneity. The identified differential DNA

methylation may also represent minor immune cell population shifts or even rare cell subtypes that are not addressed by current method. For example, a rare cell subtype in smokers that was the cause of the widely reproduced smoking-associated differential DNA methylation at *GPR15* in peripheral blood [79]. However, that does not diminish the value as a biomarkers if cost-effective but only biological interpretation of these results. More research is therefore needed to further investigate the identified pc-DMPs in terms of their presence and their stability over time in sorted blood cell types, in tumour tissues, and/or healthy tissue of the primary tumour site. This could potentially show if these are surrogate changes or due to a shared systemic response either to risk factors or cancer pathogenesis.

The pc-DMPs described in this chapter were not previously identified as age or smoking DMPs, two major risk factors for cancer, nor was there an enrichment of these risk factors in the top results. Thus, these are not biomarkers of age or smoking. BMI differences within twin-pairs may potentially contribute to cancer incidence, although only three twin-pairs comprised a cancer-affected twin that was classified as obese (BMI range 30-40 $kg/m^2$) with lower BMI measured in their healthy co-twin.

The use of a discordant MZ twin-pair design has the potential to identify changes in the DNA methylome that are independent of genetic variation or early environment. Classically, differences found within MZ twin-pairs are attributed to environmental variation. Recent publications have shown the impact of genetic variants on DNA methylation and identified for example methylation quantitative trait loci (mQTLs) and these could potentially interact with environmental exposures [203] to increase variability.

This EWAS of 41 cancer-discordant MZ twin-pairs as discovery analysis has good power to detect moderate to big effect sizes in DNA methylation, its

strength being an ultimately matched case-control study [204]. However, peripheral blood is a surrogate tissue and the heterogeneous disease might reduce this power via reduced effect sizes. The power to detect the difference in DNA methylation at the most associated pc-DMP was estimated to be 56% to reach a Bonferroni cut-off ($1.0 \times 10^{-7}$)[205]. The replication sample of nine cancer-discordant MZ twin-pairs provided 10% power to detect DNA methylation differences at the top-ranked signal at nominal significance ($p < 0.05$). Cancer discordant MZ twin-pair samples are extremely rare world-wide, however they could still provide novel indications whether similar effects are observed in an independent dataset despite the low power as shown here for the replication dataset.

## 3.5 Conclusion

This is the first EWAS for pan-cancer in peripheral blood obtained in a five-year window around time of diagnosis. Furthermore, this is a discordant MZ twin-design that is particularly powerful in detecting DNA methylation changes independent of genetic variation. One significant pan-cancer pc-DMP and three suggestive pc-DMPs as well as one pc-DMR, were identified in a sample of 41 MZ twin-pairs. Three out of the four pc-DMPs showed greater DNA methylation differences preceding cancer diagnosis and indicate regions of interest for further research into their potential as pan-cancer biomarkers.

# Chapter 4

# Early Breast Cancer Biomarkers in Discordant Monozygotic Twin-pairs

## 4.1 Background

Breast cancer is the most common cancer for women in the world with 53,354 new cases in 2013 in the UK alone [108, 206]. It comprises a heterogeneous group of tumours composed of different molecular features, prognostic behaviours, and responses to therapy [207–209]. Heritability of breast cancer was estimated close to 30% by twin studies [210, 211], though only 5-10% of cases have a strong inherited component. In familial breast cancer, a number of susceptibility genes have been identified, the most important ones being *BRCA1* and *BRCA2*. These are high-penetrance genes that have a spectrum of mutations associated with life time disease risks as high as 80% for breast cancer and 40% for ovarian cancer [212]. Multiple common genetic variants have been associated with increased breast cancer risk [213] and a subset of these have also been associated with increased risk in *BRCA1* and *BRCA2* mutation carriers [214].

Early detection is vital for optimal prognosis, and biomarkers in easily accessible tissues are being investigated that can accurately diagnose breast cancer and/or identify individuals at increased risk. The age-specific incidence rates rise steeply from the age 30-34 to age 65-69 with approximately half of the cases in the UK are diagnosed in women over the age of 65 (see Figure 4.1) [215]. Cur-

rent screening in the UK, and other economically developed countries, involves a mammography for sporadic cases between the ages of 50 to 70 and a more tailored screening protocol for high-risk individuals. Mammography screening over the last three decades has only minimally reduced the mortality rate and diagnoses of advanced state of the disease whilst unfortunately increasing the number of over-diagnosis [216]. Therefore, the need for improved and more effective screening of breast cancer remains a priority.

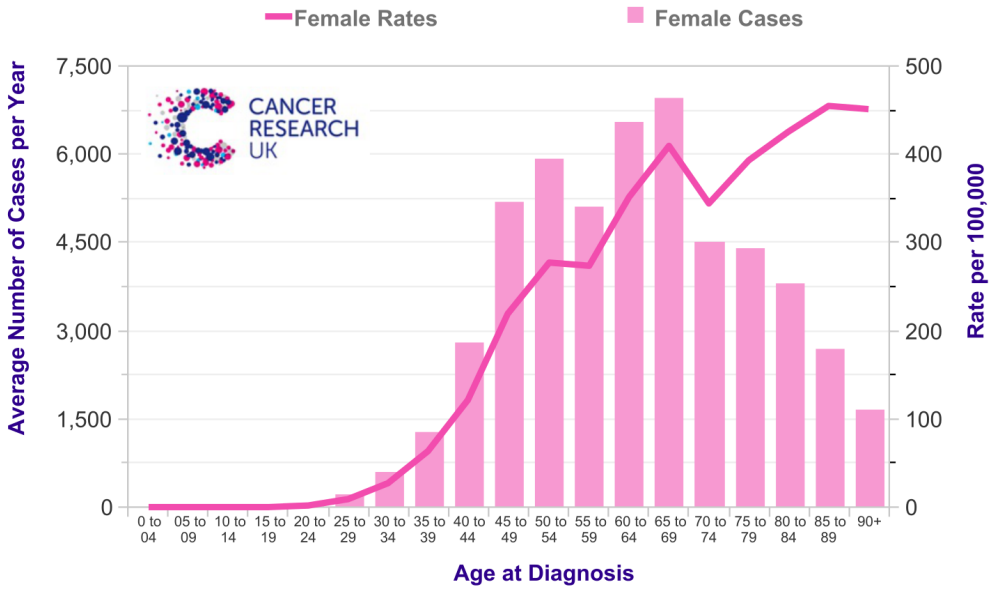



**Figure 4.1: Age-specific female breast cancer incidence rates in the UK.** Reproduced from Cancer Research UK [215].

To date, several studies have identified changes in DNA methylomes in peripheral blood samples associated with different types of cancer (discussed in section 3.1). These include a number of studies that have identified global as well as specific DNA methylation changes that are associated specifically with breast cancer [134, 137, 217–222]. In this section, three most recent studies are highlighted that identified DNA methylation changes prior to breast cancer diagnosis. Van Veldhoven *et al.* [221] identified epigenome-wide DNA hypomethylation in peripheral blood samples of breast cancer individuals compared to matched controls across three

cohorts. The DNA methylomes were assessed by the 450k for a total of 750 incident cases and matched controls. These results were validated by WGBS performed by pooling 548 DNA samples from affected individuals into four pools of DNA, as well as similar numbers for healthy matched controls. They observed that the hypomethylation in breast cancer individuals was specifically observed in gene bodies as opposed to CGIs. Yang *et al.* [222] also focused specifically on pre-diagnostic blood samples, where samples were obtained within 5 years preceding diagnosis. They compared 298 individuals diagnosed with breast cancer to 612 women who remained cancer free for 1-7 years and identified 250 differentially methylated CpGs assessed by the 27k. Thus far, one study has examined the potential of breast cancer discordant MZ twin-pairs. Heyn *et al.* [137] used 15 breast cancer discordant MZ twin-pairs from the TwinsUK cohort profiled by the 450k. They identified 403 differentially methylated CpG sites and determined via validation a candidate DMP in *DOK7* that acquires DNA methylation in the years before diagnosis in the breast cancer affected individuals.

The aim of this chapter was to investigate breast cancer associated changes in peripheral blood DNA methylomes that occur prior to diagnosis in the largest sample of breast cancer discordant MZ twin-pairs to date. Here, DNA methylomes of 28 pairs were first assessed by the 450k. To extend the DNA methylome coverage, 26 breast cancer discordant twin-pairs were also assessed by MeDIP-seq to investigate potentially all genome-wide methylated CpG sites. All blood samples were obtained within 8 years preceding breast cancer diagnosis. The most associated results, breast cancer associated DMPs (bc-DMPs) and breast cancer associated DMRs (bc-DMRs), were explored for their location in the genome and for stability over time, to reveal potential regions in the genome that can serve as breast cancer biomarkers.

## 4.2 Methods

### 4.2.1 Sample Selection

#### 4.2.1.1 Breast Cancer Discordance Criteria

Similar to chapter 3, detailed cancer diagnosis information was obtained through record linkage with the ONS. Breast cancer cases were identified by ICD-10 codes ranging from C50.1 to C50.91 and excluded C50.0 that describes malignant neoplasm of nipple and areola (see Table 4.1) [144]. Discordance was based per MZ twin-pair on one co-twin diagnosed with breast cancer within 8 years after blood sample collection whilst her co-twin was never diagnosed with malignant tumour development in the most recent records (June 2015).

**Table 4.1: ICD-10 Breast cancer codes.**

| ICD-10 code | Description |
|---|---|
| C50.1 | Malignant neoplasm of central portion of breast |
| C50.2 | Malignant neoplasm of upper-inner quadrant of breast |
| C50.3 | Malignant neoplasm of lower-inner quadrant of breast |
| C50.4 | Malignant neoplasm of upper-outer quadrant of breast |
| C50.5 | Malignant neoplasm of lower-outer quadrant of breast |
| C50.6 | Malignant neoplasm of axillary tail of breast |
| C50.8 | Malignant neoplasm of overlapping sites of breast |
| C50.9 | Malignant neoplasm of breast of unspecified site |

#### 4.2.1.2 Sample Selection Per Platform

Breast cancer discordant MZ twin-pairs were selected from the TwinsUK registry (see Figure 4.2) for wich peripheral blood DNA methylomes were profiled by the 450k and/or by MeDIP-seq.

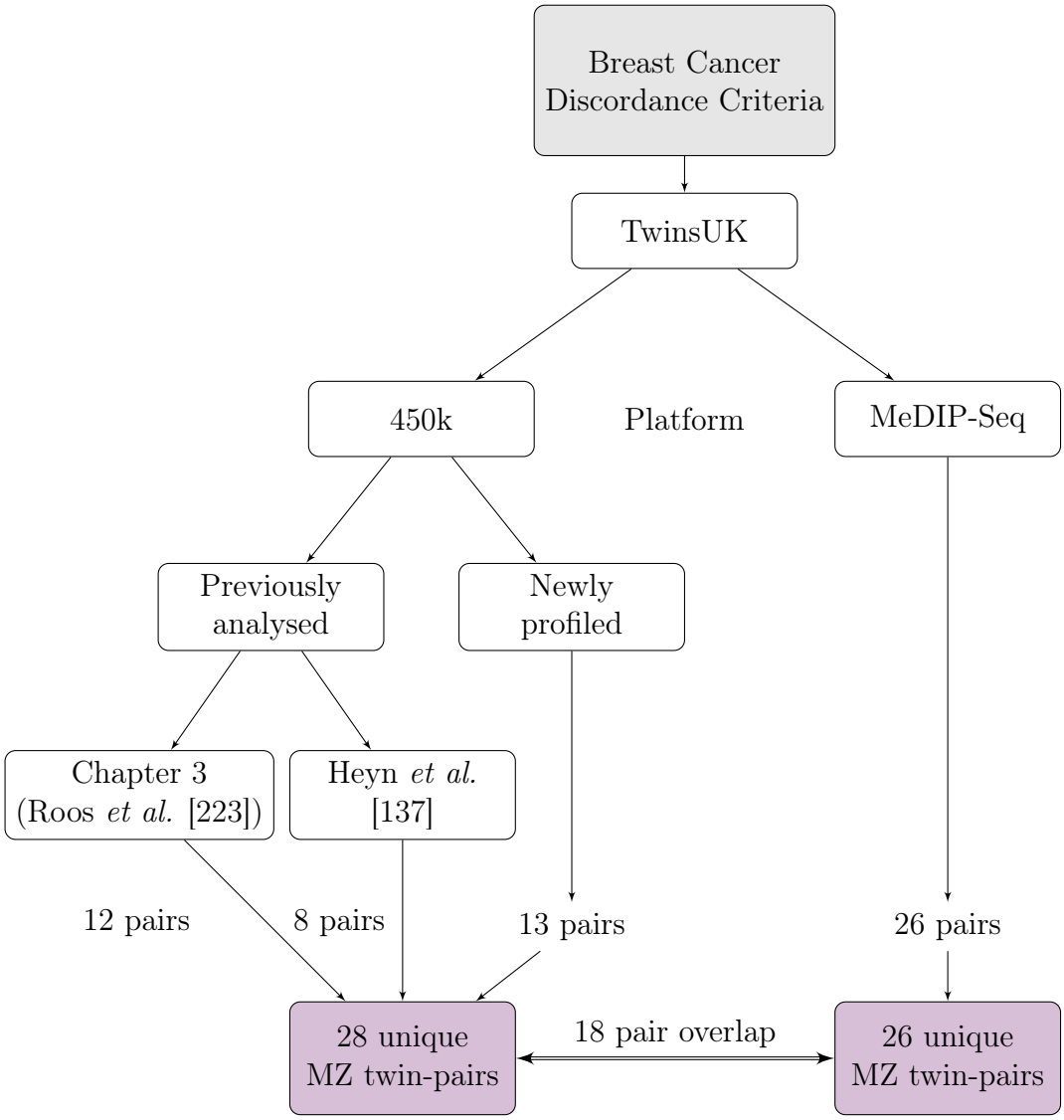



**Figure 4.2: Schematic overview of sample selection.** The numbers shown in this flowchart represent the total number of MZ twin-pairs after extensive quality control that were included in the downstream analyses.

**Selected Samples Profiled By the 450k**

In total, 28 middle-aged female discordant breast cancer MZ twin-pairs of European descent were selected, ranging from 21 to 78 years old at blood collection (see Table 4.2). The co-twins not diagnosed with cancer were at least cancer free in a period ranging from 2.6 to 17 years following diagnosis of her co-twin (median = 7.5 years). Of these 28 MZ twin-pairs, 13 were newly profiled at IDIBELL and had not been included in published data to date. The remaining 15 unique MZ twin-pairs included 12 pairs profiled at the Wellcome Trust Sanger Institute, previously described in chapter 3 (9 pairs were included in the pan-cancer main analysis), and 8 MZ twin-pairs profiled at IDIBELL that were part of the samples published in Heyn *et al.* [137] (see Table 4.3). There is an overlap of 6 MZ twin-pairs that were aliquots from the same blood samples and subsequently profiled at both genomic centres, leaving 15 unique MZ twin-pairs across these centres.

**Table 4.2: Characteristics of 28 breast cancer-discordant MZ twin-pairs.**

| Selection | Characteristic | Mean | Median | Range | |
|---|---|---|---|---|---|
| 28 discordant twin-pairs | Age at DNA extraction | 58.5 | 59.3 | 20.6 | 78.7 |
| | Age at Cancer diagnosis | 60.9 | 62.4 | 23.3 | 82.6 |
| | Cancer Free (yrs) | 8.7 | 7.5 | 2.6 | 17 |
| 26 discordant twin-pairs | BMI* | 2.2 | 1.6 | 0.1 | 9.2 |
| 54 Individuals | BMI | 27.0 | 26.6 | 18.4 | 38.3 |

* In absolute differences.

**Table 4.3: Distribution over genomic centres of 28 breast cancer discordant MZ twin-pairs.**

| | IDIBELL (Barcelona) | Wellcome Trust Sanger Institute (Hinxton) | Unique MZ twin-pairs |
|---|---|---|---|
| Unpublished number of pairs | 13 | 0 | 13 |
| Published number of pairs | 8 | 12* | 15* |
| Total number of pairs | 22 | 12 | 28 |

*3 MZ twin-pairs not included in main analyses.

Major risk factors for cancer pathogenesis, smoking and BMI, were investigated in the 28 MZ twin-pairs. Smoking habits were assessed from longitudinal questionnaires and individuals were divided into three categories: never smoked, current smokers, and ex-smokers that stopped at least 3 years before blood sample collection (see Table 4.4). 19 MZ twin-pairs were concordant in smoking habit: 9 pairs were non-smokers, 1 pair were current smokers, and 9 pairs were ex-smokers. The remaining 9 twin-pairs comprised 6 pairs including an ex-smoker and never smoker co-twin and 3 pairs including an ex-smoker and current smoker co-twin. For 26 MZ twin-pairs BMI was measured during a clinical visit. The mean BMI of all individuals was 26.9 kg/m$^2$ and in 10 twin-pairs, the twin diagnosed with cancer had a greater BMI compared to her co-twin. The mean and median absolute within-pair differences were relatively small, 2.2 and 1.6 kg/m$^2$ respectively. BMI was not associated with cancer status in these individuals ($p = 0.82$).

**Table 4.4: Smoking habits of 28 breast cancer-discordant MZ twin-pairs.**

| MZ twin-pair | Never smoked | Ex-smoker | Current smoker | Total |
|---|---|---|---|---|
| Concordant (number of pairs) | 9 | 9 | 1 | 19 |
| Discordant (number of pairs) | 6 | 3 | 8 | 9 |

### Selected Samples Profiled by MeDIP-seq

This selection included 26 middle-aged female discordant breast cancer MZ twin-pairs, ranging from 23 to 79 years old at blood collection (median and mean of 58 years) of European descent (see Table 4.5). Of these, 18 MZ twin-pairs have also been processed with the 450k from the same blood sample. The co-twins not diagnosed with cancer, were cancer free in a period ranging from 2.6 to 21 years following diagnosis of her co-twin. All these DNA methylomes have not been published in a cancer related analysis before.

The major risk factors for cancer pathogenesis, smoking and BMI, were also

assessed in these 26 MZ twin-pairs. 16 MZ twin-pairs were concordant in smoking habit: 10 pairs were non-smokers, 1 pair were current smokers, and 5 pairs were ex-smokers (see Table 4.6). The remaining 10 twin-pairs comprised 7 pairs including an ex-smoker and never smoker co-twin, 2 pairs including an ex-smoker and current smoker co-twin, and one pair comprised a current smoker and never smoker. The mean and median BMI of all individuals was 26.5 kg/m$^2$ and 26.6 kg/m$^2$ respectively. In 16 twin-pairs, the twin diagnosed with cancer had a lower BMI compared to her co-twin. The mean and median absolute within-pair differences were similar to the 450k sample, 1.9 and 1.4 kg/m$^2$ respectively. BMI was again not associated with cancer status in these individuals ($p = 0.68$).

**Table 4.5: Characteristics of 26 breast cancer-discordant MZ twin-pairs.**

| Selection | Characteristic | Mean | Median | Range | |
|---|---|---|---|---|---|
| 26 discordant twin-pairs | Age at DNA extraction | 59.1 | 60.0 | 20.6 | 78.7 |
| | Age at Cancer diagnosis | 61.7 | 62.9 | 23.3 | 82.6 |
| | Cancer Free (yrs) | 8.9 | 6.7 | 2.6 | 20.6 |
| | BMI* | 2.0 | 1.5 | 0.1 | 7.8 |
| 52 Individuals | BMI | 26.3 | 26.5 | 18.4 | 35.2 |

* In absolute differences.

**Table 4.6: Smoking habits of 26 breast cancer-discordant MZ twin-pairs.**

| MZ twin-pair | Never smoked | Ex-smoker | Current smoker | Total |
|---|---|---|---|---|
| Concordant (number of pairs) | 10 | 5 | 1 | 16 |
| Discordant (number of pairs) | 7 | 2 | 1* | 10 |

* Comprised a current smoker and never smoker.

## 4.2.2 Genome-wide DNA Methylation Data

### 4.2.2.1 Illumina 450k

Peripheral blood DNA methylomes profiled with the 450k from bisulphite-converted DNA were pooled across two genome centres; IDIBELL (two batches of

8 and 14 samples) and the Wellcome Trust Sanger Institute (two batches of 2 and 10 samples). The array, pre-processing, and quality control are described in detail in section 2.3.1.

In short, probes were removed that: 1) failed detection in one or more samples and/or had a bead count less than 3 in >5% of samples (n=2,297), 2) aligned to more than one location in the human genome with their 50 bp sequence, 3) located on the sex chromosomes, 4) harboured common genetic variants occurring in European Caucasians (MAF >1%) within 10 bp on the probe at the interrogated CpG site, and 5) contained variants at any MAF at the interrogated CpG site [150, 151]. The remaining number of probes that wer included in the main EWAS was 431,673.

Furthermore, the 28 MZ twin-pairs were verified with the sample identifier (see Section 2.3.1.4) using the 57 autosomal control SNP probes and known genotype data. The beta values were then normalized using functional normalisation that removes technical variation using control probes [155].

Cell type proportions were estimated using the method devised by Houseman *et al.* [80], for CD8+ T cells, CD4+ T cells, B cells, Natural Killer cells, granulocytes, and monocytes. Pairwise correlations between these proportions were subsequently assessed by Spearman's Rank correlation and revealed that granulocytes were strongly correlated with all other cell types except monocytes (see Figure 4.3).

PCA was performed on normalised beta values (N(0,1) per probe. The first four PCs combined explained 43.3% of the total variance in DNA methylation. No nominally significant association was observed between cancer status and the first four PCs. Significant ($p$ <0.01) associations with the first four PCs included: Granulocytes, CD8+ T cells, Natural Killer cells, B cells, family ID, bead chip, as well as centre and batch.

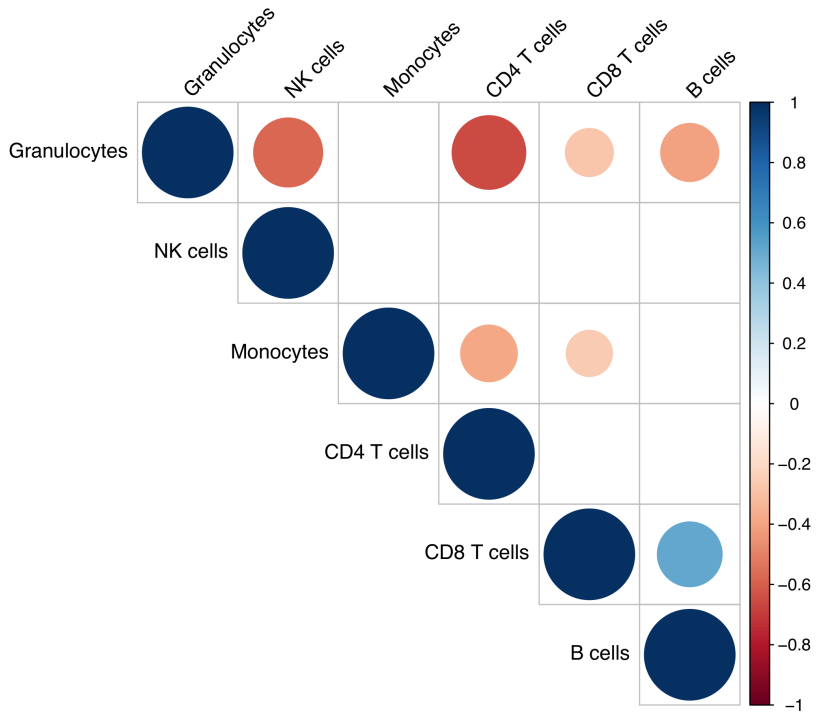



**Figure 4.3: Pair-wise correlations of estimated cell type proportions.** The size of the circles represents the strength of the Spearman's Rank correlation. Only correlations are shown with a $p$ value $<0.01$.

#### 4.2.2.2 MeDIP-seq

Peripheral blood DNA methylomes of the 26 MZ twin-pairs were profiled with MeDIP-seq at BGI, Shenzen, China. The blood samples were profiled in six batches of 8, 4, 25, 3, 8, and 4 samples. Apart from one MZ twin-pair, both twins were included in the same batch. The 26 MZ twin-pairs on average had a coverage of ~17 million unique mapped reads and were selected for further quality control. The methodology and initial QC applied in this sample is described in detail in section 2.3.2.

As described in section 2.3.2.2, bins comprised 500 bp sliding windows with 250 bp overlap covering the entire genome. The total number of atuosomal bins was 11,524,145 across the genome. These were further investigated for regions of no general coverage, defined here as zero reads. Per bin, the threshold of zero reads was

set at 10% of individuals, *i.e.* no more than 5 individuals were allowed to have zero reads per bin. This reduced the total number bins of to 5,055,534 for downstream analysis. Although the method of MeDIP-seq infers that zero reads could indicate a completely unmethylated region, true hypomethylated functional regions generally still have low levels of methylation as seen in single CpG bisulphite sequencing DNA methylome studies [10, 41]. Therefore regions with no reads are likely to be false positive hypomethylated regions, taking also into account the coverage of uniquely mapped reads. In this analysis of a relatively small number of individuals, the random occurrence of more zero reads in either the case or control group can have a strong impact. This restriction of no general coverage thus prevent that potential findings will be driven by the (technical) artifact of low coverage.

Measured cell counts were available for a subset of 20 MZ twin-pairs, however within this subset, 11 pairs these were not from the same date of blood sample collection.

### 4.2.3 Statistical Analysis

A general overview of the analyses performed in this chapter is shown in Figure 4.4.

#### Global DNA Methylation Profiles

Genome-wide DNA methylation variation was analysed using unsupervised hierarchical clustering analysis with Euclidean distances and complete linkage method on the 450k samples.

#### Breast Cancer DMPs

The DNA methylomes were first adjusted for covariates using a linear model fitted on standardised beta values per probe (N(0,1)) and the estimated cell type proportions of granulocytes and monocytes as well as centre and batch (four levels). The residuals from this model were then used to calculate within twin-pair differences

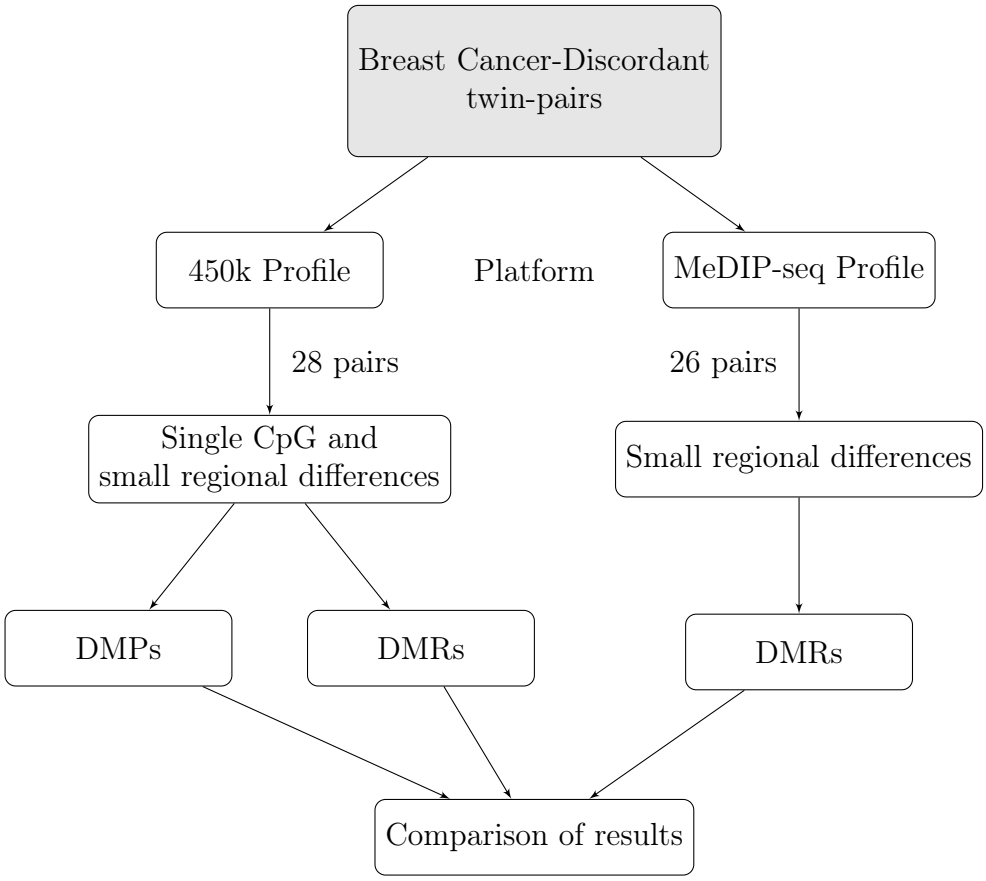



**Figure 4.4: Schematic overview of statistical analyses.** A general overview of the analyses performed in this chapter.

that were determined consistently as the residual of the cancer-affected twin minus the residual of the healthy co-twin. A one-sample t-test on these differences was then performed to assess significance. Epigenome-wide significance level was adjusted for multiple testing by use of a FDR of 10% using "qvalue" in R and a suggestive results threshold was set at a nominal $p$ value of 1 x $10^{-5}$.

### Breast Cancer DMRs by Illumina 450k

An EWAS was performed at small genomic regions to identify bc-DMRs similar to described in 3.2.4 for pc-DMRs. Here, small regions were predefined and included at least 2 CpG sites no more than 500 bp apart. To keep the paired structure of the data, differences at single CpG sites within these regions were per MZ twin-pair determined on DNA methylation residuals, similar to the DMP analysis. This was

then compared to a group without DNA methylation differences using the "Bumphunter" package in R [176]. The algorithm determines a $p$ value based here on 1,000 permutations as well as determining FWER adjusted $p$ value. Bc-DMRs were identified with a FWER adjusted $p$ value $<0.05$.

**Breast Cancer DMRs by MeDIP-Seq**

To identify bc-DMRs, within twin-pair differences were determined consistently as cancer-affected twin minus healthy co-twin and a one-sample t-test was performed on these data. Three different models were explored preceding the calculation of within twin-pair differences: 1) standardised (N(0,1)) RPM, 2) standardised (N(0,1)) RPM adjusted by batch in a linear model, and 3) for a subset of 20 MZ twin-pairs standardised (N(0,1)) RPM adjusted by batch and measured cell counts (lymphocytes, neutrophils, monocytes, and eosinophils) in a linear model. The models are compared in the results section (see Section 4.3.4) based on the genomic inflation factor, $\lambda$, of the Q-Q plots using "GenABEL" in R [224].

The main results described in this chapter were based on the differences of unadjusted standardised (N(0,1)) RPM. Significance levels were adjusted for multiple testing by use of a FDR of 5% using "qvalue" in R while suggestive results threshold was set at a nominal $p$ value of $1 \times 10^{-7}$.

### 4.2.4 Genomic Annotation Analysis

Annotation of CpG sites and regions was peformed with respect to CpG density (CGI, shores, and shelves), relative to RefSeq genes (promoter, 5'UTR end, gene body, 3'UTR, intergenic), and functional genomic elements derived from ENCODE including ChromHMM state segmentation, DNase-I hypersensitivity sites, and TFBSs [178, 179].

## 4.3 Results

### 4.3.1 Peripheral Blood DNA Methylome Profiles

Peripheral blood DNA methylomes were analysed for global differences in 28 female breast cancer discordant MZ twin-pairs interrogated with the 450k. The individuals were diagnosed with breast cancer only and provided blood samples for analysis that were obtained up to 8 years preceding diagnosis. Unsupervised hierarchical clustering of unadjusted normalised DNA methylomes was performed to assess global variation between all individuals (see Figure 4.5). Individuals in all but one twin-pair clustered together as pairs (96.4%), and thus did not show global differences according to breast cancer status. There is a clear batch cluster for the samples analysed at IDIBELL. Therefore, downstream analysis of the 450k DNA methylome data included an adjustment for batch effects.

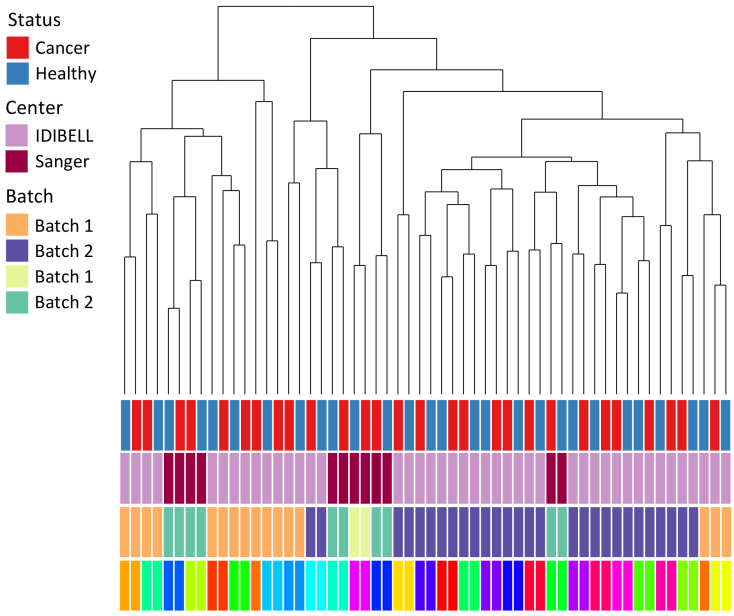



**Figure 4.5: Dendrogram of 56 whole blood DNA methylomes.** Annotation bars coloured for affection status per individual, centre, batch per centre, as well as family identifier (see legend).

### 4.3.2 Breast Cancer Associated DMPs

DNA methylation levels at single CpG sites were analysed epigenome-wide within the 28 female breast cancer discordant MZ twin-pairs to identify bc-DMPs. Preceding the EWAS, the DNA methylomes were adjusted for estimated cell type proportions of granulocytes and monocytes as well as centre and batch. The EWAS was then performed by a one-sample t-test on the directional adjusted DNA methylation differences within twin-pairs, that is breast cancer twin minus healthy co-twin. No epigenome-wide significant bc-DMPs were identified, however, four novel bc-DMPs were identified at a suggestive threshold of $p$ <1.0 x $10^{-5}$ (see Figure 4.6 and Table 4.7 on page 109).

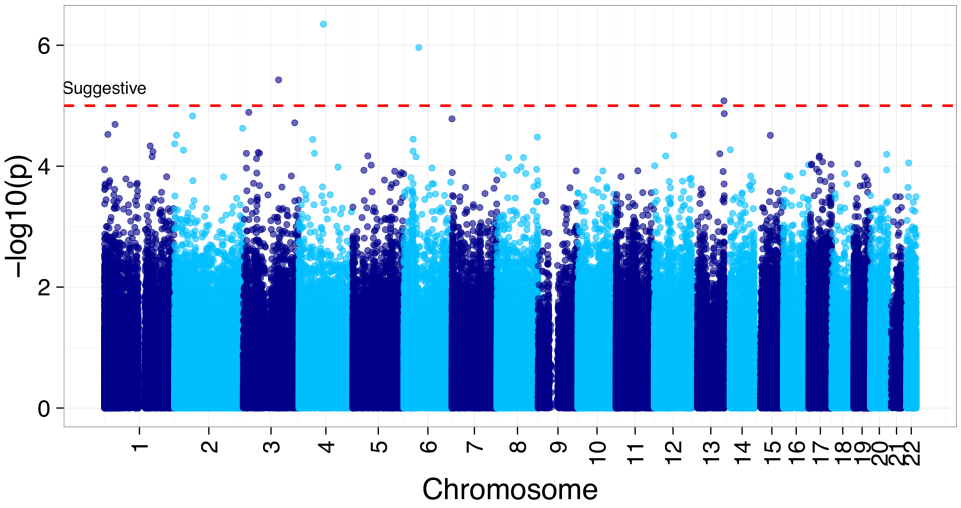



**Figure 4.6: Manhattan plot of breast cancer EWAS results.** Each point represents the observed $-\log_{10}$ $p$ value at a single CpG-site. The suggestive threshold of 1 x $10^{-5}$ is shown as a red striped vertical line.

The two most associated bc-DMPs passed a FDR threshold of 20%. These included cg21446955 in *ARHGAP24* ($p$ = 4.5 x $10^{-7}$) and cg04165000 in *GCLC* ($p$ = 1.1 x $10^{-6}$). Cg21446955 resides right at the start of the 5' UTR of *ARHGAP24* and showed consistently lower DNA methylation levels in breast cancer-affected twins compared to their healthy co-twins. The directional difference observed in the normalised unadjusted DNA methylation had a similar direction of effect with

a mean of -1.5% DNA methylation difference with a range of -3.9% to 1.0% (see Figure 4.7). Opposite the effect observed here, knockdown of *ARHGAP24* in triple negative breast cancer cells results in enhanced invasion capacity [225] and shows attenuated expression via *ETS2* in breast cells carrying a mutation in *p53* [226]. Cg04165000 is located in the body of *GCLC* that also showed consistently lower DNA methylation levels in breast cancer-affected twins compared to their healthy co-twins (mean of -1.5% in unadjusted DNA methylation within a range of -10.8% and 3%, see Figure 4.7). *GCLC* is an essential enzyme for the tripeptide Glutathione (GSH), one of the key players in the antioxidant defenses of the cell [227]. GSH is required for cancer initiation and high levels are observed in many tumours to increase the antioxidant capacity [227, 228].

A further two suggestive associations were determined for the CpG sites interrogated by probes cg21069563 at the 5' UTR of *SLC41A3* ($p = 3.7 \times 10^{-6}$) and cg25947845 in the gene body of *MCF2L* ($p = 8.3 \times 10^{-6}$). A similar direction of effect was observed at both CpG sites of lower DNA methylation levels in breast cancer-affected individuals compared to their healthy co-twins (see Figure 4.7). GWAS results have identified a SNP in *MCF2L* to be associated with increased risk of bladder cancer [229] and higher expression of *MCF2L* has been observed across different tumours [230]. Furthermore, gene body hypomethylation has been identified in prostate cancer and in the surrounding tissue (adjacent and distant) compared to tissue of healthy volunteers [231].

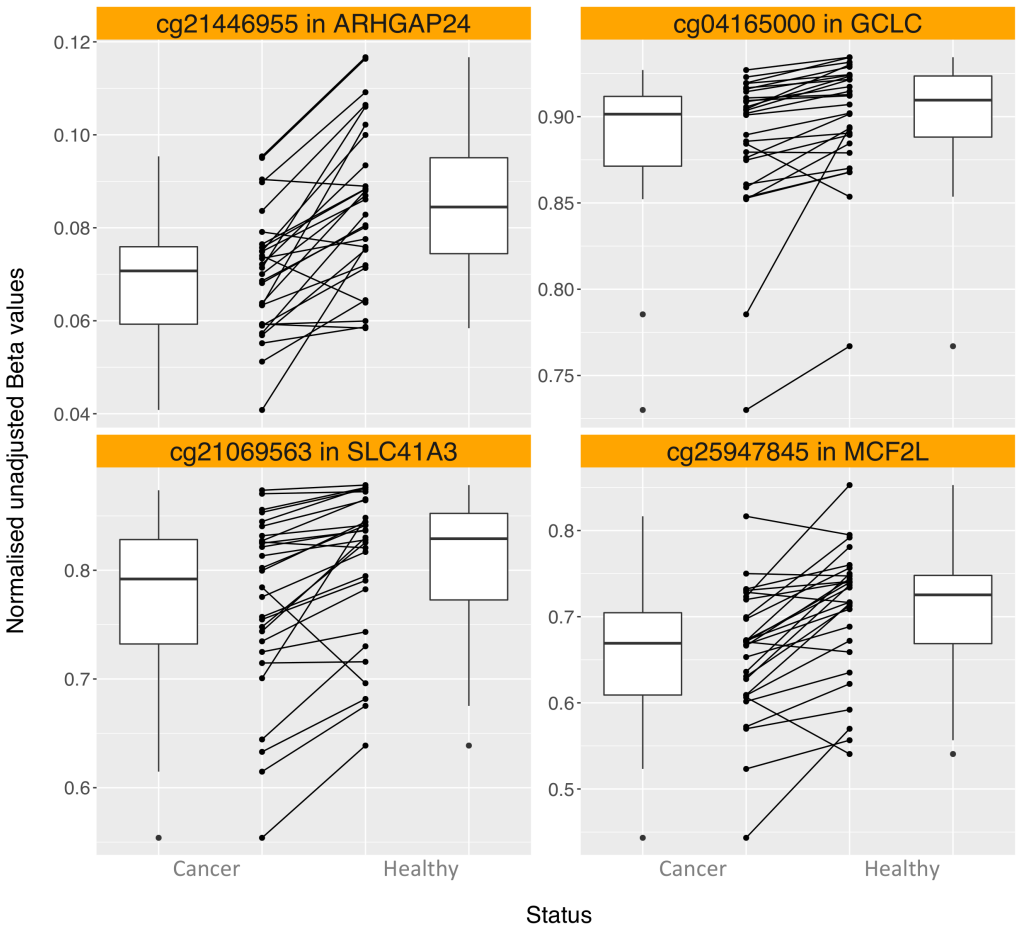



**Figure 4.7: Associations between MZ twin-pairs at four suggestive breast cancer DMPs.** Normalised unadjusted beta values are shown of cancer-affected individuals (left) and healthy individuals (right) with the lines connecting each MZ twin-pair.

**Table 4.7: Four top-ranked DMPs from EWAS of 28 breast cancer discordant MZ twin-pairs.**

| Rank | CpG | Position (hg19) | Gene | Location | CpG density | Mean difference* | *P* value |
|---|---|---|---|---|---|---|---|
| 1 | cg21446955 | chr4:86,851,425 | *ARHGAP24* | 5' UTR | - | -0.86 | 4.5 x $10^{-7}$ |
| 2 | cg04165000 | chr6:53,375,279 | *GCLC* | Body | - | -0.55 | 1.1 x $10^{-6}$ |
| 3 | cg21069563 | chr3:125,800,198 | *SLC41A3* | 5'UTR | Shelf | -0.53 | 3.7 x $10^{-6}$ |
| 4 | cg25947845 | chr13:113,718,045 | *MCF2L* | Body | Shelf | -0.66 | 8.3 x $10^{-6}$ |

*The mean differences are calculated as breast cancer - unaffected co-twin using adjusted DNA methylation values.

**Table 4.8: Four top-ranked DMRs from EWAS of breast cancer 28 discordant MZ twin-pairs.**

| Rank | Position (hg19) | Gene | Location | CpG density | Number of CpGs | Direction | *P* value | FWER | Direction CpG sites |
|---|---|---|---|---|---|---|---|---|---|
| 1 | chr4:698,302-698,615 | *PCGF3* | TSS 1500 | Island | 3 | - | 8.3 x $10^{-6}$ | 0.024 | - + - |
| 2 | chr7:73,417,196-73,417,263 | - | - | - | 3 | - | 1.4 x $10^{-5}$ | 0.037 | - - - |
| 3 | chr1:39,571,502-39,571,504 | *MACF1* | Body | Island | 2 | - | 2.5 x $10^{-5}$ | 0.055 | - - |
| 4 | chr5:39,425,137-39,425,449 | *DAB2* | TSS 200 - 1st Exon | Island - Shore | 4 | - | 5.2 x $1^{-5}$ | 0.059 | - - - - |

### 4.3.3 Breast Cancer Associated DMRs

Next, DNA methylation levels at small genomic regions were assessed for association with breast cancer status in the 28 MZ twin-pairs to identify bc-DMRs. Similar to methods described in chapter 3, regions were predefined to contain at least two CpGs no more than 500 bp apart. At each CpG site within these regions, the directional within-pair difference in adjusted DNA methylation was determined using the same adjusted DNA methylation data as described in the previous section 4.3.2. Four bc-DMRs were identified after 1,000 permutations with "Bumpunter" [176] (FWER <0.1, see Table 4.8). The two most associated bc-DMRs were identified at a epigenome-wide significant FWER of <0.05.

The most associated bc-DMR overlaps a CGI ~1,500 bp from the TSS of *PCGF3* ($p$ = 8.3 x $10^{-6}$, see Figure 4.9 A). The region is annotated as weakly transcribed in the GM12878 cell line by ENCODE. It encompasses three single CpG sites that show hypomethylation in breast cancer affected twins at two CpGs that were independently nominally significant ($p$ <0.05) (see Figure 4.8). *PCGF3* belongs to the polycomb group proteins that when deregulated, contribute to the pathogenesis of multiple cancers [232, 233].

The second ranked bc-DMR ($p$ = 1.4 x $10^{-5}$) was identified ~20 kb upstream of its nearest gene *ELN* (see Figure 4.9 B). Located in a region annotated as weak/poised enhancer in GM12878 by ENCODE, it spans three CpG sites that are all hypomethylated in breast cancer affected twins (see Figure 4.8).

Two more bc-DMRs were identified in *MACF1* ($p$ = 2.5 x $10^{-5}$, see Figure 4.9 C) and *DAB2* ($p$ = 5.2 x $10^{-5}$, see Figure 4.9 D). The first lies in a CGI within the gene body of *MACF1* and is again hypomethylated in breast cancer affected twins (see Figure 4.8). The region is marked by active promoter states in all ENCODE cell lines and spans two CpG sites. Somatic mutations in *MACF1* itself have been observed in multiple cancers, including breast [234], and has been indicated as a marker for survival prognosis [235]. The second is located over the

TSS of *DAB2* and partly overlapped a CGI. Except for the GM12878 cell line, it lies within an active promoter (ENCODE) and is hypomethylated in breast cancer affected twins (see Figure 4.8). *DAB2* is a well known tumour suppressor gene [236–238] and therefore does not directly link biologically with the observed decreased DNA methylation at its TSS in this study.

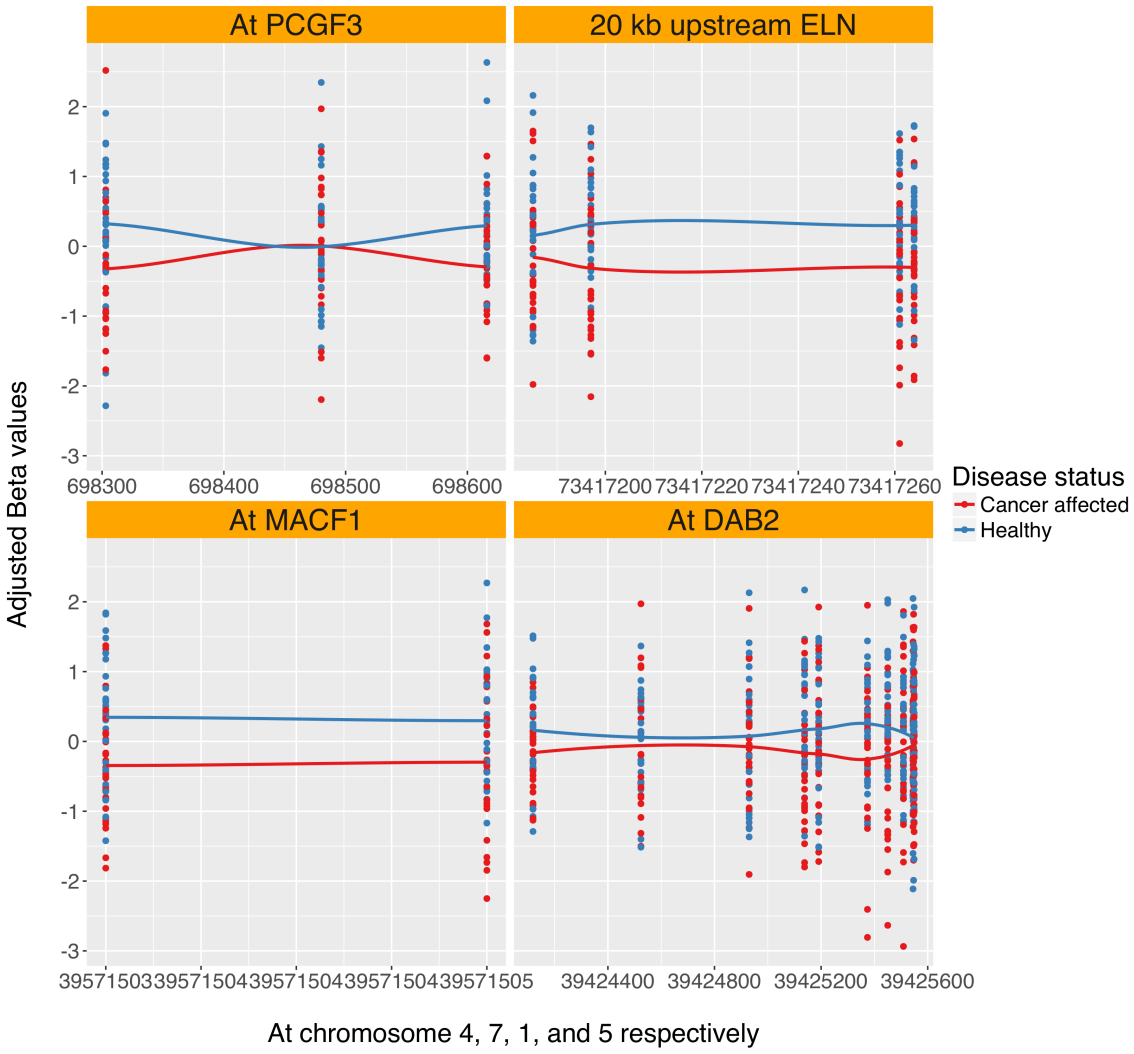



**Figure 4.8: Associations between MZ twin-pairs at four breast cancer DMRs.** Adjusted DNA methylation values at each CpG site in the DMR and smooth (LOESS) lines are shown for individuals affected by breast cancer (red) and healthy co-twins (blue).

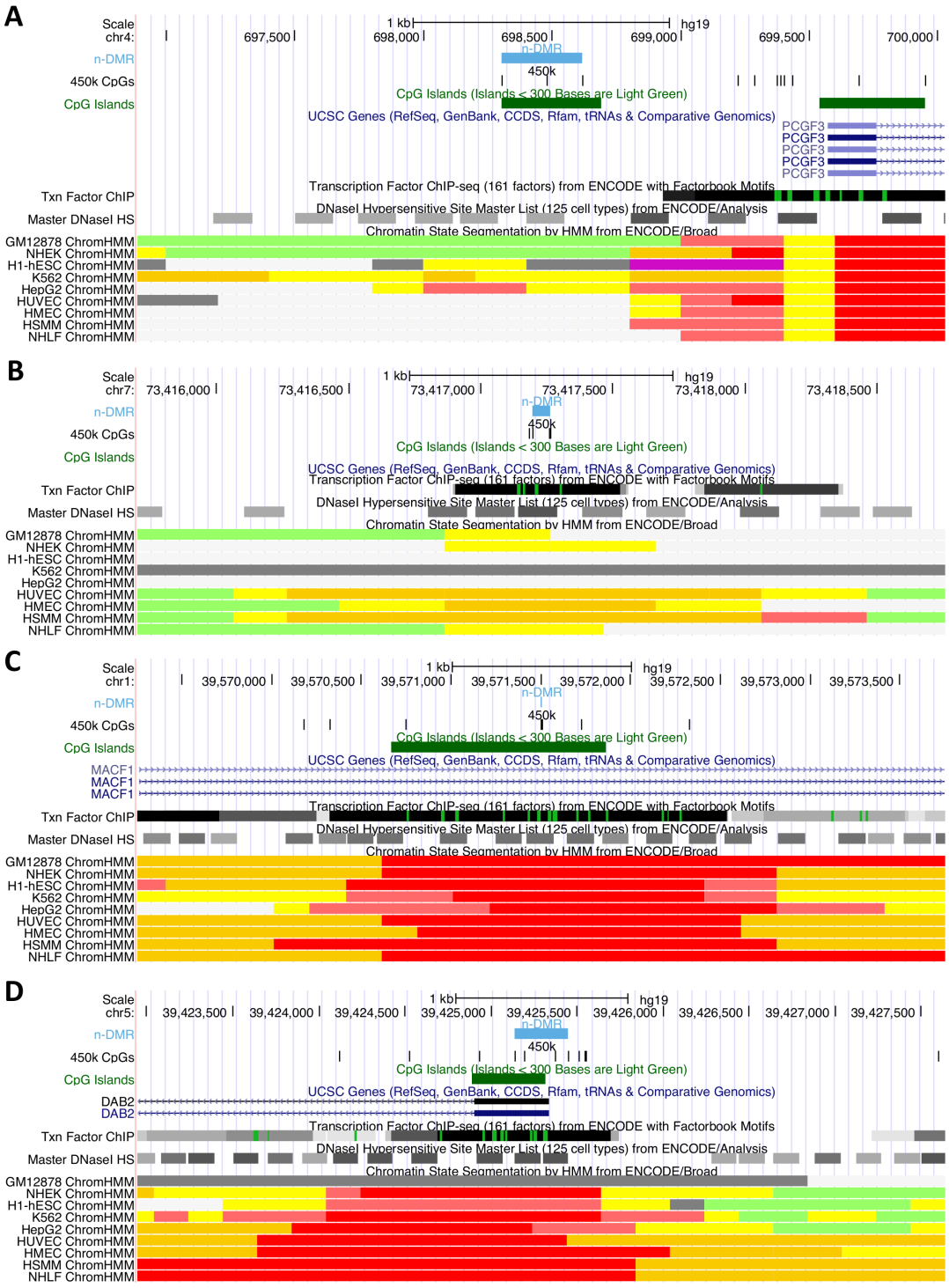



**Figure 4.9: Location of the four breast cancer DMRs in the human genome.** Figures obtained from UCSC Genome browser [239], displaying position in the genome (hg19), CpG sites from the 450k, n-DMR (in light blue), RefSeq genes, CGI, transcription factor ChIP data, DNase-I sensitivity sites, and ChromHMM genomic segmentation. **(A)** At *PCGF3*. **(B)** ~20 kb upstream *ELN*. **(C)** At *MACF1*. **(D)** At *DAB2*.

### 4.3.4 Breast Cancer Associated DMRs By MeDIP-seq

Regional DNA methylation differences were investigated in 26 breast cancer discordant MZ twin-pairs but now using a different method to interrogate the DNA methylomes, MeDIP-seq. Again, a one-sample t-test on the directional differences within MZ twin-pairs was performed. Preceding that, three different steps were performed on standardised methylation values (RPM) per bin; 1) No adjustment, 2) Adjusted for batch effects, and 3) Adjusted for batch effects and blood cell counts of lymphocytes, neutrophils, monocytes, and eosinophils (in a subset of 20 MZ pairs). The Q-Q plots of the observed -$\log_{10}$ $p$ values for all three workflows are shown in Figure 4.10. The genomic inflation factor, λ, was used to assess systemic bias characterized by the extent to which the –$\log_{10}$ $p$ values deviated from the expected uniform distribution [240]. The standardised unadjusted DNA methylation analysis results show the least systemic bias with a λnear 1. The other two models have smaller λ's and appear to over-correct demonstrated by the $p$-values below the expected uniform distribution. When $p$ values and ranking order was compared across the three models, the most associated bins were very similar (see Figure 4.11). Therefore the main results focus on the model with no adjustment prior to the discordance EWAS on standardised methylation values per bin.

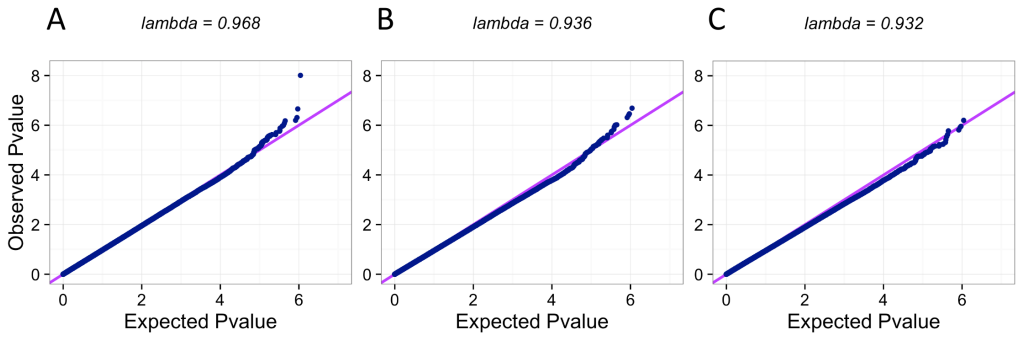



**Figure 4.10: Q-Q plot of observed –$\log_{10}$ $p$ values against the expected –$\log_{10}$ $p$ values from the expected distribution per breast cancer EWAS. (A)** No adjustment, **(B)** Adjusted for batch, and **(C)** A subset of 20 pairs adjusted for batch and blood cell counts of lymphocytes, neutrophils, monocytes, and eosinophils.

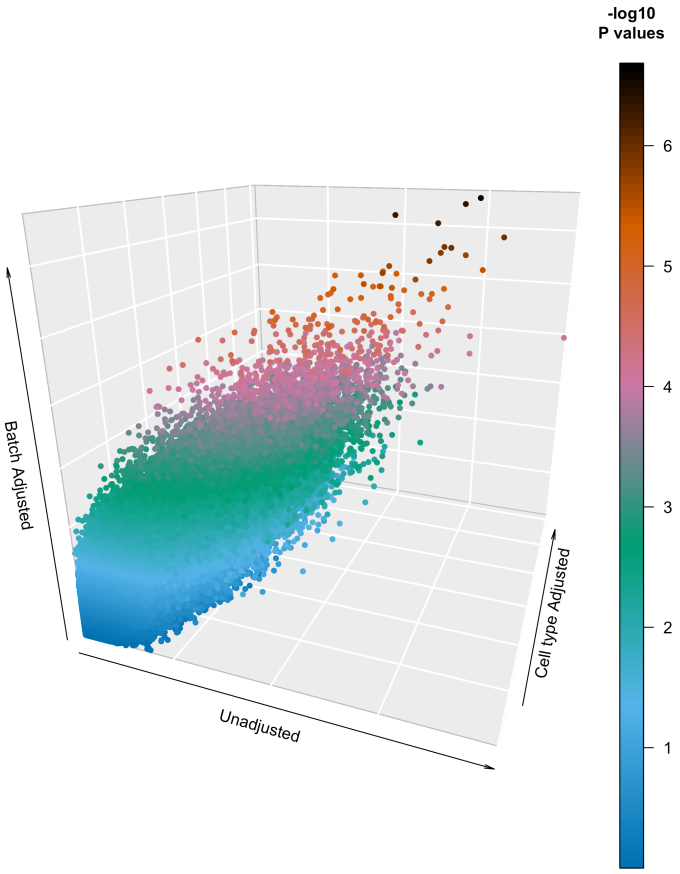



**Figure 4.11: Three dimensional plot of observed $-\log_{10}p$ values from three EWASs.** Each point depicts -$\log_{10}$ $p$ value of the three models.

One novel bc-DMR was identified in *MECOM* (*MDS1* and *EVI1* complex locus) at a FDR of 5% (chr3:168,887,751-168,888,250, $p = 9.8 \times 10^{-9}$, see Figure 4.12 and Figure 4.13). DNA methylation levels at this locus were consistently higher in the breast cancer affected twins compared to their healthy co-twins. The RPM scores at this locus have a median directional difference of 0.16 within a range of -0.09 to 0.30 RPM (see Figure 4.14 A). The relationship was subsequently assessed between date of blood sample collection and date of breast cancer diagnosis, revealing a consistent pattern of difference spanning the 8 years to diagnosis (see Figure 4.14 B). This bc-DMR also remained significant in the two other models correcting for batch and measured cell counts; $p = 7.1 \times 10^{-5}$ and $p = 3.8 \times 10^{-6}$ respectively.

Common genetic variation in *MECOM* predisposes individuals to myeloproliferative neoplasms [241] and somatic copy number variations (CNVs) of *MECOM* have been long established in, but not limited to, ovarian cancers [242–245]. Furthermore, isoforms of *MECOM* are implicated in transcription regulation across the genome and the oncogenicity of ovarian cancer [246] as well as being upregulated itself in metastatic breast cancer [247].

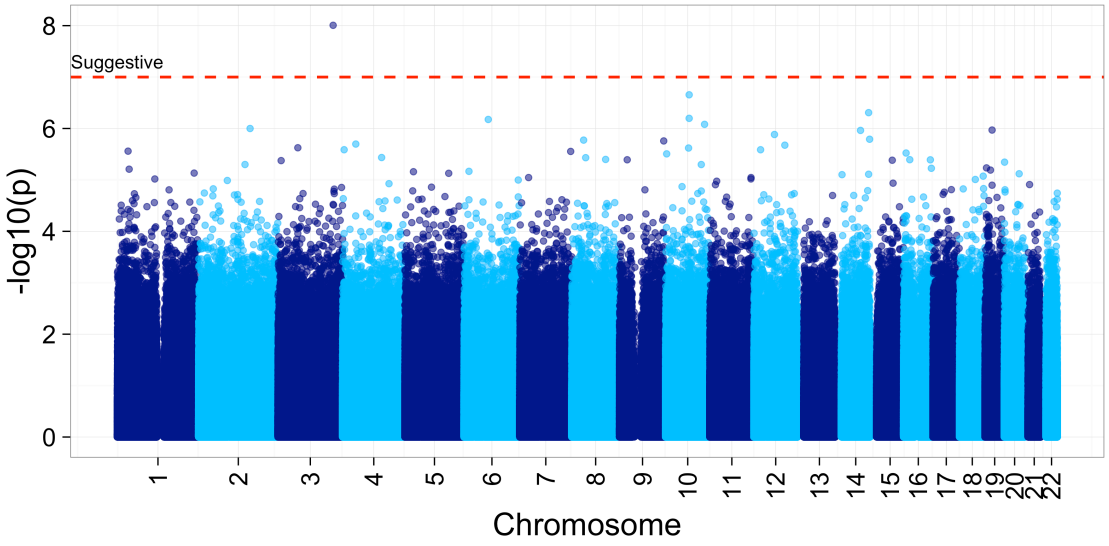



**Figure 4.12: Manhattan plot of breast cancer EWAS by MeDIP-Seq.** Each point represents the observed $-\log_{10}$ $p$ value at a single bin. A suggestive threshold of 1 x $10^{-7}$ is shown as a red striped vertical line.

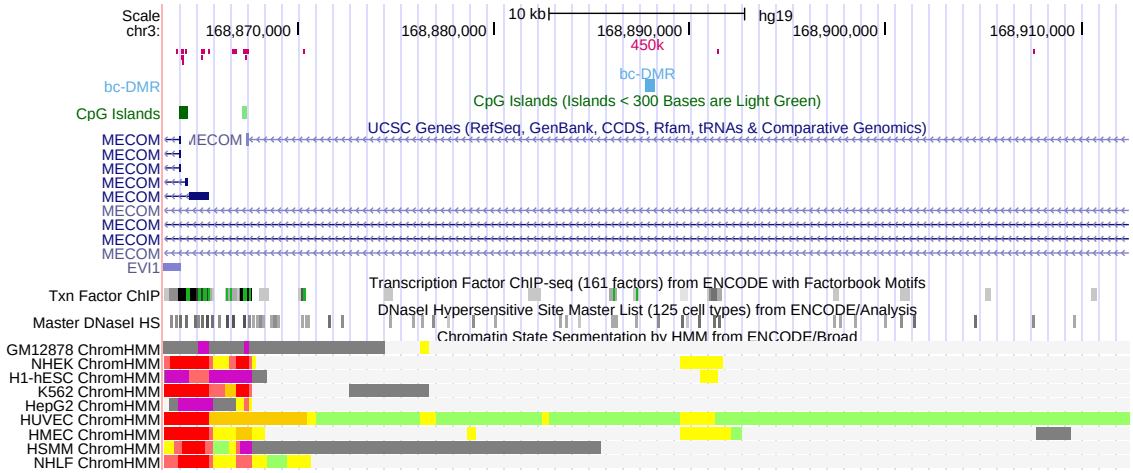



**Figure 4.13: Location of the breast cancer DMR in the human genome.** Figures obtained from UCSC Genome browser [239], displaying position in the genome (hg19), CpG sites from the 450k, bc-DMR (in light blue), RefSeq genes, CGI, transcription factor ChIP data, DNase-I sensitivity sites, and ChromHMM genomic segmentation.

Associations between MZ twin-pairs at four breast cancer DMRs.

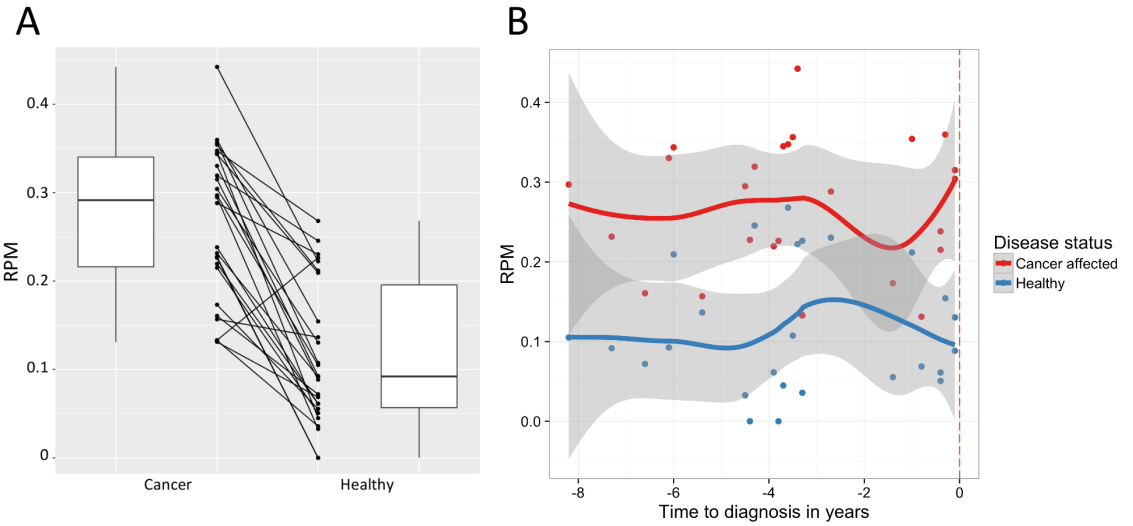



**Figure 4.14: Association between MZ twin-pairs at the breast cancer DMR in *MEDCOM*. (A)** RPM are shown of cancer-affected individuals (left) and healthy individuals (right) with the lines connecting each MZ twin-pair. **(B)** RPM are shown of each individual according to the time the blood sample was obtained prior to diagnosis and are coloured by disease status (see legend).

### 4.3.5 Comparison of Results By Illumina 450k and MeDIP-seq

Finally, the top ranked results of both methods were compared. Analysis on DNA methylation interrogated by MeDIP-seq identified one bc-DMR in *MECOM* that passed a FDR of 5%. Interrogation by 450k identified four bc-DMRs at *PCGF3*, ~20 kb upstream of *ELN*, *MACF1*, and *DAB2* (FWER <0.1) as well as four suggestive bc-DMPs at *ARHGAP24*, *GCLC*, *SLC41A3*, and *MCF2L* ($p$ <1.0 x $10^{-5}$). Both datasets and top results were overlapped with a window of 1 kb to include adjacent bins or probes.

Increased genome-wide coverage is one of the strengths of MeDIP-seq and accordingly the bc-DMR identified with MeDIP-seq is not profiled by the 450k. The nearest probe is an isolated single probe with no others in its vicinity and lies ~3 kb downstream. It was was not significantly associated with breast cancer in the single CpG site EWAS ($p = 0.8$).

Of the four bc-DMRs of the 450k analyses, only the top ranked bc-DMR at *PCGF3* overlapped bins available in MeDIP-seq data. This bc-DMR overlapped two bins and also had three adjacent bins within 1 kb that all showed the same direction of association with breast cancer that is, lower DNA methylation in the breast cancer twin, but no nominal significance level was reached. The remaining bc-DMRs at ~20 kb upstream of *ELN*, *MACF1*, and *DAB2* had 7, 3, and 6 bins available in the MeDIP-seq data respectively. The vast majority showed a similar (negative) direction of effect as observed in the 450k, except two bins on the edge of the 1 kb windows. None of these bins in the MeDIP-seq data reached nominal signifance. Near nominal significance was observed at the bc-DMR ~20 kb upstream of *ELN* with a bin ~450 bp downstream ($p = 0.058$) and a bin ~250 bp upstream ($p = 0.078$).

The four most associated bc-DMPs of the 450k analyses were next considered. Apart from the top ranked result at *ARHGAP24*, the other three bc-DMPs

overlapped at least one bin available in MeDIP-seq data. The differential DNA methylation findings on both platforms showed the same direction of association with breast cancer, that is, lower DNA methylation in the breast cancer twin, but did not reach nominal significance level in the MeDIP-seq dataset. At the bc-DMP in *ARHGAP24*, the nearest MeDIP-seq bin was 76 bp downstream of the 450k CpG site. All were also hypomethylated in the breast cancer twin but again no nominal significance level was reached with the strongest $p$ value of 0.14.

## 4.4 Discussion

Here, the largest sample to date of 28 breast cancer discordant MZ twin-pairs were analysed for differential DNA methylation changes in peripheral blood samples. DNA methylomes were assessed using two methods; 450k (28 pairs) and MeDIP-seq (26 pairs). One epigenome-wide significant bc-DMR (<FDR 5%) in *MECOM* was identified via MeDIP-seq. This region showed a consistent pattern of higher DNA methylation levels in breast cancer affected individuals compared to their healthy co-twins from 8 years to year of diagnosis. Four bc-DMRs in *PCGF3*, ~20 kb upstream *ELN* (at <FWER 5%), *MACF1*, and *DAB2* (at <FWER 10%) were identified via the 450k regional method analysis. The single CpG site EWAS identified four suggestive bc-DMPs in or near *ARHGAP24*, *GCLC*, *SLC41A3*, and *MCF2L* ($p$ <1.0 x $10^{-5}$). Globally the MZ twin-pairs were more similar, with respect to breast cancer affection status.

The novel identified bc-DMR in *MECOM* by MeDIP-seq had consistently greater DNA methylation levels in the twin with cancer than the healthy co-twin. In RPM, this is a difference of 0.16 within a range of -0.09 to 0.30 that is consistent across the years to diagnosis. This region was also significant in the two other EWASs correcting for batch and measured blood cell counts. The signal lies in the first intron and upstream of *MECOM* isoform variants in a region that is annotated as low signal heterochromatin in all cell lines of ENCODE except in HUVEC, umbilical vein endothelial cell line, where it is annotated as a weakly transcribed region. It is closely flanked by DNase-I sensitive sites as well as near various TFBSs. This region is not targeted by the 450k and demonstrates the strength and need for approaches with greater coverage than that available provided by the 450k and other beadchips. *MECOM* itself is implicated various cancers. Common genetic variation in *MECOM* increases the risk of myeloproliferative neoplasms [241] and somatic CNVs in this gene that have been identified in ovarian cancers and other

cancers [242–245]. In particular, isoforms of *MECOM* as well as the alternatively spliced gene *EVI1* located on the same strand within the transcript of *MECOM*, are implicated in differential transcription and contribute to pathogenesis of ovarian cancers [246] as well as metastatic breast cancer [247]. This bc-DMR within the intron of *MECOM* could therefore contribute to driving certain isoforms that could aid cancer pathogenesis.

Four small genomic regions were also identified via the 450k through a regional method analysis. Of those, the most strongly associated bc-DMR overlaps a CGI ~1,500 bp from the TSS of *PCGF3*. The region is annotated as weakly transcribed in GM12878 by ENCODE and exhibits consistent lower DNA methylation levels in breast cancer affected twins compared to their healthy co-twins. This region also showed the same direction of association in the two MeDIP-seq bins that overlapped it, though these bins did not reach significance. *PCGF3* is part of the polycomb group proteins that are epigenetic regulators of transcription and function in polycomb repressive complexes (PRCs) that can modify histones and act to silence genes [248]. When deregulated, these contribute to the development of various cancers [232, 233].

Another bc-DMR was identified in *MACF1* overlapping a CGI in the first intron and upstream isoform variants. It is located in a region that is annotated as active promoter in all ENCODE cell lines and exhibits lower DNA methylation in breast cancer affected twins compared to their healthy co-twins. A similar direction of association was observed in two nearest bins in the MeDIP-seq dataset. Somatic mutations in this gene have been identified in breast cancer among multiple other cancers [234] and are also markers of survival prognosis [235].

The single CpG site analysis of the 28 breast cancer discordant MZ twin-pairs identified four suggestive results that did not reach epigenome-wide signifi-

cance. The most strongly associated bc-DMP was at the start of the 5' UTR of *ARHGAP24* and had consistently lower DNA methylation levels in the twin with cancer than the healthy co-twin. This was also observed in the nearest MeDIP-seq bin 76 bp downstream, although no signficance was reached. The region is annotated by ENCODE as low transcribed heterochromatin in GM12878 cell lines, but also as active promoter in HepG2 (hepatocellular carcinoma cell line) and as a weak enhancer in other cell lines such as epidermis, skeletal muscle, lung fibroblasts, and embryonic stem cells (NHEK, HSMM, NHLF, and H1-hESC) indicating a region that is different across tissues. Knockdown of *ARHGAP24* in triple negative breast cancer cells is associated with enhanced invasion capacity [225]. Furthermore, it also shows reduced expression in breast cells with a *p53* mutation via transcription factor *ETS2* [226].

The bc-DMP in the fifth intron of *GCLC* showed lower DNA methylation values in the breast cancer affected twin compared to their healthy co-twins. MeDIP-seq bins adjacent this bc-DMP showed a similar pattern. This CpG site lies in a region annotated by ENCODE as weakly transcribed as well as transcriptional elongation. *GCLC* is an essential enzyme for the tripeptide GSH, one of the pivotal players in the antioxidant defense system of the cell [227]. Higher levels of GSH are observed in many tumours and required for cancer initiation due to the toxic conditions in tumours [227, 228].

This study used two different approaches of interrogating DNA methylomes, the widely used 450k and MeDIP-seq that was used for the first time in breast cancer biomarker analysis. In essence, MeDIP-seq is a whole genome approach and uses an antibody against DNA methylation to pull down methylated DNA fragments for next generation sequencing that results in a regional quantification of DNA methylation. The 450k on the other hand is a targeted approach to a set of predetermined CpG sites that uses bisulphite treatment to identify DNA

methylation that results in a single CpG site quantification of DNA methylation. A comparison between the two by Clark *et al.* [249] determined an autosomal correlation of DNA methylation values of 0.68. In differential methylation detection by both methods however, there was not a high overlap of results; >64% of DMPs identified by the 450k were covered by MeDIP-seq but not identified as DMRs and vice versa although percentages are not clearly stated in the publication.

When the results of both sets of analyses described in this chapter were compared there was little to no overlap in the differentially methylated sites and regions. Partly because the bc-DMR in *MECOM* did not have probes on the 450k within its vicinity, and only one of the four bc-DMRs from the 450k data were covered by reads in the MeDIP-seq data. When the bins adjacent to the 450k bc-DMRs within 1 kb were included, the majority of bins the direction of association was similar for the two technologies but the MeDIP-seq results did not reach nominal significance. Three of the four bc-DMP were directly covered by reads in the MeDIP-seq data. Again a similar direction of association was observed in bins overlapping and near to the bc-DMPs.

This lack of direct overlap of significance was not unexpected and could reflect these two different methods of quantification of DNA methylation that was also observed by Clark *et al.* [249]. Nevertheless, the vast majority of adjacent and overlapping bins to the 450k results did show a similar direction of association.

In the MeDIP-seq data there were several quality control and analysis decisions that could impact the downstream results. One of these was based on the exclusion of genomic regions where a very low number of unique reads was obtained across individuals. The number of aligned reads suggested to cover the majority of the methylated CpGs in the genome (80% of total CpGs) with at least one read is ~60 million [250]. Here, the coverage per sample is less with an estimate of ~17 million reads, covering close to 60% of the total CpGs with at least one read from the

estimated curve of Taiwo *et al.* [250]. As a consequence, bins with no reads could not be robustly interpreted as completely unmethylated and most likely would have arisen from no coverage. The exclusion of these bins more importantly minimised results that were driven only by the strong effects of MZ twin-pairs where one twin had in fact no coverage whilst the co-twin had coverage. This conservative threshold ensured only bins were analysed that were adequately covered by MeDIP-seq. This has no impact on the bins included in this analysis, however, it could have excluded bins that could have been true signals from the analysis.

The main MeDIP-seq results discussed here where from the model with no adjustment prior to the discordance EWAS on standardised methylation values per bin. This data was not further adjusted for technical effects and blood cell composition as was performed for the 450k data. The reasoning of using the unadjusted standardised RPM was determined by the use of genomic inflation factor, $\lambda$, to measure systemic bias in the three EWAS pipelines. The use of $\lambda$ may not be ideal for measuring systemic bias, as an EWAS is not similar to GWAS, nevertheless it is most used at this point in time to evaluate models. The analysis on unadjusted standardised DNA methylation levels was also compared to analyses adjusted for batch and for batch and measured cell counts (in a subset of 20 MZ twin-pairs). The top ranked results were comparable across all three models. The lower $\lambda$ for batch adjusted values could be due the amount of levels (6) that for a subset of bins does not perform well. The lower $p$ values in the cell type corrected data could reflect to the lower power due to a smaller sample size and because not all measured cell counts were available from the same time point as the blood sample.

A more general limitation of the study is the complex heterogeneity of breast cancer with varying pathogenesis. This in theory can dilute the power to detect differences prior to diagnosis due to different aetiology of the subclasses of breast cancer that might be captured differently in DNA methylomes. As mentioned

in the previous chapter, potential general systemic effects or surrogate effects observed in blood could occur across individuals [168–172]. By using MZ twin-pairs we minimised the genetic variation known to have strong effects that influence the quantification of DNA methylation by the 450k and more so in the MeDIP-seq dataset. Here, the observed differences in DNA methylation may potentially be more prone to environmental effects due to the exclusion of genetic differences.

## 4.5 Conclusion

This study has used the largest sample to date of breast cancer discordant MZ twin-pair peripheral blood DNA methylomes and targets the time window before diagnosis. The analyses included both MeDIP-seq and 450k technologies to assay methylomes, and therefore ensure a reasonable genome coverage of the breast cancer peripheral blood DNA methylome using a combination of methylation profiling approaches. Three epigenome-wide significant novel bc-DMRs were identified in *MECOM*, *PCGF3*, and ~20 kb upstream of *ELN*. Furthermore, two suggestive bc-DMRs and four suggestive bc-DMPs in *MACF1*, *DAB2*, *ARHGAP24*, *GCLC*, *SLC41A3*, and *MCF2L* respectively, were identified. In all cases the signals are of interest from a translational point of view as they are present before diagnosis and are relevant for future prognostic and diagnostic studies. Further research into these signals and their replication, could lead to the assessment of their potential to be used as blood based biomarkers for breast cancer.

# Chapter 5

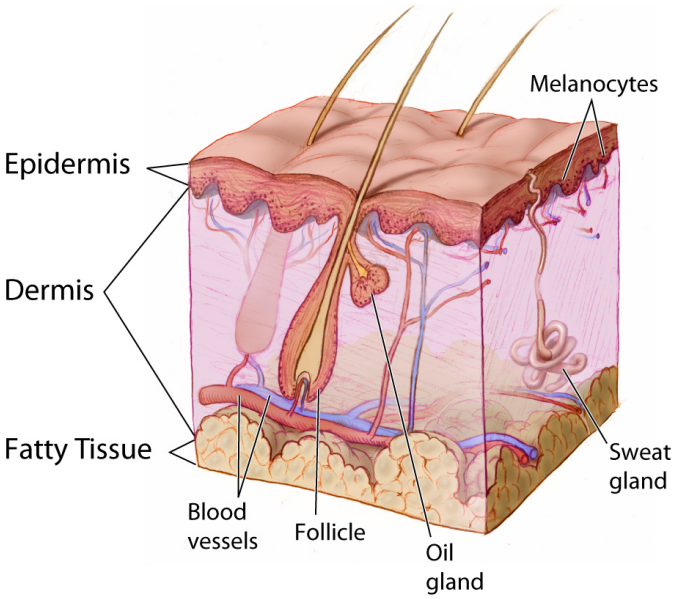
# Higher Naevus Count Exhibits A Distinct DNA Methylation Signature in Healthy Human Skin

## 5.1 Background

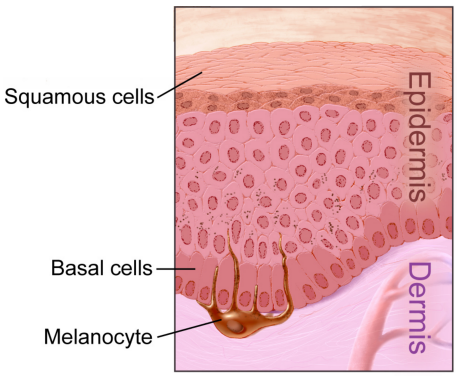
Melanocytic naevi, commonly called moles, are benign lesions comprised of a clonal proliferation of melanocytes. They are more common in light-skinned populations [251]. The total number of melanocytic naevi[1] counted across the entire body is the strongest known risk factor for melanoma in Caucasian populations [252, 253]. Melanoma is the ninth most diagnosed cancer in Europe with over 100,000 new cases in 2012, which amounts to 3% of all cancer cases (excluding non-melanoma skin cancers) and its incidence rates are increasing [254, 255]. It is the third most prevalent cancer of the skin and the most aggressive of these. In 20 to 40% of cases melanoma arise from existing benign naevi, with the remaining majority arising from new melanocytic lesions [256–258]. Naevi are vastly more common than melanomas and therefore the number of naevi act as a marker of risk, considering that the majority of naevi do not progress to melanoma [256]. New insights into the biology of naevi and the predisposition factors that influence the skin's propensity for melanocyte proliferation, observed as the occurrence of increased number of naevi, will ultimately improve the understanding of melanoma pathogenesis.

[1] In the remainder of this thesis this is shortened to "naevi" meaning melanocytic naevi.

Human skin contains three defined tissue layers, the upper two layers comprising the epidermis and dermis, and a deeper subcutaneous layer consisting of fat and connective tissue, the hypodermis (see Figure 5.1 A). As the name suggests, melanomas arise from melanocytes. These are melanin producing cells originating from neuronal crest cells (melanoblasts). Generally melanocytes reside in the basal layer of the epidermis with a ratio of one melanocyte to every 10 keratinocytes (see Figure 5.1 B) [259]. Clonal proliferation of melanocytes gives rise to naevi which can be congenital and occurs in approximately 1% of newborns [260, 261]. The vast majority of naevi however, is acquired after birth. These naevi show distinct histopathologic features [262] and are labelled as benign tumours of melanocytes.




**5.1(A)** Healthy skin




**5.1(B)** Melanocyte location

**Figure 5.1: Simplified representation of healthy skin. (A)** Three layers of healthy human skin. **(B)** Zoomed in image of epidermis and dermis depicting the location of majority of melanocytes. Both images were released by the National Cancer Institute, part of the National Institutes of Health. Created by Don Bliss (Illustrator).

Typically, naevi are acquired from birth until approximately 40 years of age, in particular during early childhood, adolescence, and pregnancy, and thereafter decrease in number [263]. Men and women have different patterns of distribution of naevi across the body in Caucasian populations, with women having more naevi on the arms and legs, whereas men tend to have more on the trunk [264, 265]. The reason for this gender difference is still unclear and disputed. Proposed hypotheses include differences in sun exposure habits or differences in melanocyte differentiation in early embryogenesis between males and females [266]. There is also a corresponding gender-specific difference observed in the location of the majority of melanomas: on the trunk for men and on the legs for women [267].

The decrease in number of naevi after the age of 40 is delayed in individuals at high risk of melanoma, who therefore display an altered senescence of naevi [268]. Higher numbers of naevi have been associated with longer telomere length in peripheral blood [269], as well as reduced (or lack of) sun damage as represented by the occurance of solar keratoses [265, 270, 271]. This may indicate a difference in senescence pathways between individuals, which is reflected in the number of naevi and possibly could be detected in the skin itself.

A genetic basis for the number of naevi across the body has been demonstrated by two GWASs with five associated SNPs to date, located in loci at *PLA2G6* (2), *MTAP*, *NID1*, and near *C11orf74* [272, 273]. Moreover, the SNPs in *PLA2G6* have been replicated in these two separate studies. Total body naevus count has also been shown to be a useful intermediate phenotype for melanoma as the same SNPs have been identified to be associated with both traits at *PLA2G6* and *MTAP* [272–275]. The phenotype variance explained by these SNPs however is low, as observed for many complex traits studied in GWAS. The role of epigenetic variation associated with the number of naevi has not been explored yet in healthy skin or other tissues.

In this chapter, DNA methylation variation in healthy human skin tissue was investigated with total body naevus count for 322 female individuals. Variation was explored at both individual CpG sites for total body naevus count associated DMPs (n-DMPs), as well as small regions for total body naevus count associated DMRs (n-DMRs). The DNA methylation results were further examined for gene expression changes in the same skin tissue. Finally, GWAS SNPs associated with naevus count or melanoma risk were analysed for influence on DNA methylation levels in *cis*.

## 5.2 Methods

### 5.2.1 Sample Selection

Individuals were selected from female twins from the TwinsUK, for whom skin tissue DNA methylome data was available (468 samples). Trained dermatology research nurses performed detailed total body naevi counts following a standardised and reproducible naevus count protocol as described previously [276]. Total body naevus count was the sum of all naevi $>$ 2 mm across all sites of the body. In total, 322 females were selected of which the skin biopsy samples were obtained on average 9.7 years after examination within a range 6.5-11.9 years (see Table 5.1 and Figure 5.2 A). Individuals were excluded that were diagnosed to date with cancers of the skin. In total, this sample included 25 MZ twin-pairs, 65 DZ twin-pairs, and 147 unrelated individuals.

**Table 5.1: Characteristics of 322 female individuals.**

| Characteristic | Mean | Median | Range | |
|---|---|---|---|---|
| Age | 59.4 | 60.6 | 38.7 | 83.1 |
| BMI | 26.6 | 26 | 16.2 | 47.1 |
| Naevus count | 34.1 | 19 | 0 | 231 |
| Time between biopsy and naevus count | 9.7 | 9.8 | 6.5 | 11.9 |

These individuals were aged 39 to 83 years with a median of 60.6 years. Smoking habits were assessed at time of the skin biopsy, similar to previous chapters, from longitudinal questionnaires and divided into three categories: never smoked, current smokers, and ex-smokers (stopped $<$3 years before skin biopsy). This resulted in 159 individuals that had never smoked, 38 individuals were current smokers, and 126 individuals were ex-smokers. Lastly, BMI ranged from 16.2 to 47.1 kg/m$^2$ with a median of 26 kg/m$^2$ (see Figure 5.2 B).

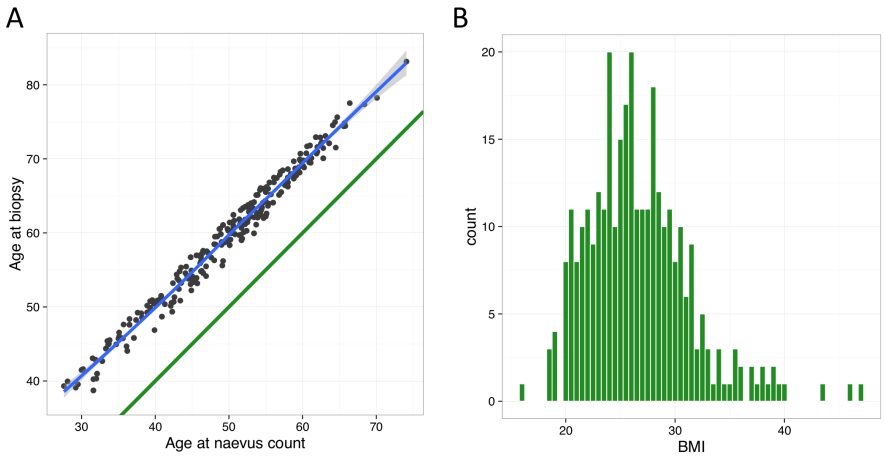



**Figure 5.2: Age and BMI at naevus examination. (A)** Correlation of chronological ages between naevus examination and skin punch biopsy. The blue line is the true least squares regression fit whereas the green line is a perfect age match. **(B)** BMI at naevus examination.

### 5.2.2 Genome-wide DNA Methylation Profiles

Skin DNA methylomes were profiled with the 450k from bisulphite-converted DNA extracted from skin punch biopsies that were not co-located with visible naevi, in two batches consisting of 82 and 245 samples. The punch biopsies (8 mm) were taken from an area adjacent and inferior to the umbilicus directly followed by mechanically dissection of the fat layer (hypodermis) before snap-freezing. The beadchip, pre-processing, and QC are described in detail in section 2.3.1.

In short, probes were removed if they failed detection in at least one sample and with a bead count less than 3 in more than 1% of the samples (n = 18,208), if the 50 bp sequence aligned to multiple locations in the genome, if the probes were located on the sex chromosomes, and if the probes contained SNPs at MAF >1% within 10 bp of the interrogated CpG site or at any MAF at the interrogated CpG site. Therefore the remaining number of probes used for the EWAS was 415,909. All probes containing the SNP exclusion criteria were included for the GWAS SNP regional figures and are highlighted as such in figures and accompanying text.

All individuals were verified with the sample identifier (see section 2.3.1.4) using the 57 autosomal control SNP probes and known genotype data. Based on low mean overall intensity signals (combined unmetylated and methylated values), four individuals were excluded and the remaining 322 samples were normalised using the BMIQ method to correct for probe type bias [154].

PCA on epigenome-wide profiles of the 322 individuals was performed using standardised beta values (N(0,1)) per probe. The first 3 PCs combined explained 36% of the total variance in these DNA methylomes and these PCs were assessed for associations with likely confounders that included: beadchip, position on the beadchip, age, smoking status, BMI, and bisulphite conversion efficiency (as measured by the 450k control probes). Strong associations ($p < 1 \text{ x } 10^{-20}$) were identified with beadchip and bisulphite conversion efficiency.

### 5.2.3 Gene Expression Profiles

Gene expression profiles were obtained for a subset of 248 individuals from the same skin tissue biopsies which were also profiled for DNA methylation as part of the MuTHER project [146] described in detail in section 2.4. These profiles were previously normalised with quantile normalization of the three replicates of each individual followed by quantile normalization across all individuals.

This subset of 248 individuals had similar characteristics in age, BMI, and proportionally similar distribution in smoking habits. The age range was 39 to 83 years of age with a median of 60.5 years. 127 individuals that had never smoked, 26 individuals were current smokers, and 95 individuals were ex-smokers. BMI ranged from 16.2 to 47.1 kg/m$^2$ with a median of 25.9 kg/m$^2$. This subset comprised 14 MZ twin-pairs, 40 DZ twin-pairs, and 140 unrelated individuals. Slightly higher proportions were observed of unrelated individuals (56% compared to 45%) and DZ twin-pairs (32% compared to 25%) to the complete set of 322 individuals.

### 5.2.4 Genotypes

Genotype data were obtained for a subset of 283 individuals of Caucasian ancestry from the TwinsUK described in detail in section 2.5. In short, imputation was performed with IMPUTE using the 1000 Genomes data phase 3 reference panel. Quality control genotype measures included thresholds for minimum genotyping rate (>95%), Hardy–Weinberg equilibrium ($p$ >1.0 x $10^{-6}$), and MAF (>1%). The imputation quality score was >0.5 for GWAS catalogue SNPs.

### 5.2.5 External Datasets

A publicly available epidermal and dermal DNA methylome dataset profiled with the 450k from Vandiver *et al.* [277] was downloaded from the gene expression omnibus (GEO) database with the accession code "GSE51954". This comprised 20 unique individuals that underwent two skin biopsies each. These were then mechanically separated into epidermis and dermis, resulting in a total of 40 dermal and 38 epidermal DNA methylomes. VanDiver *et al.* described two categories of 10 "younger" individuals (<35 years of age) and 10 "older" individuals (>60 years of age). These two categories were also divided in sun exposed and sun protected sites (see Table 5.2). Location of the biopsies was from the upper inner arm (n = 20) for the sun protected sites and either the dorsal fore arm (n = 5) or lateral to the eye (n = 5) for the sun exposed sites.

**Table 5.2: Epidermal and dermal DNA methylome characteristics.**

| **Age category** | **Epidermis** | | **Dermis** | |
|---|---|---|---|---|
| | Sun exposed | Sun protected | Sun exposed | Sun protected |
| Younger individuals | 9 | 9 | 10 | 10 |
| Older individuals | 10 | 10 | 10 | 10 |

### 5.2.6 Statistical Analysis

A general overview of the analyses performed in this chapter is shown in Figure 5.3.

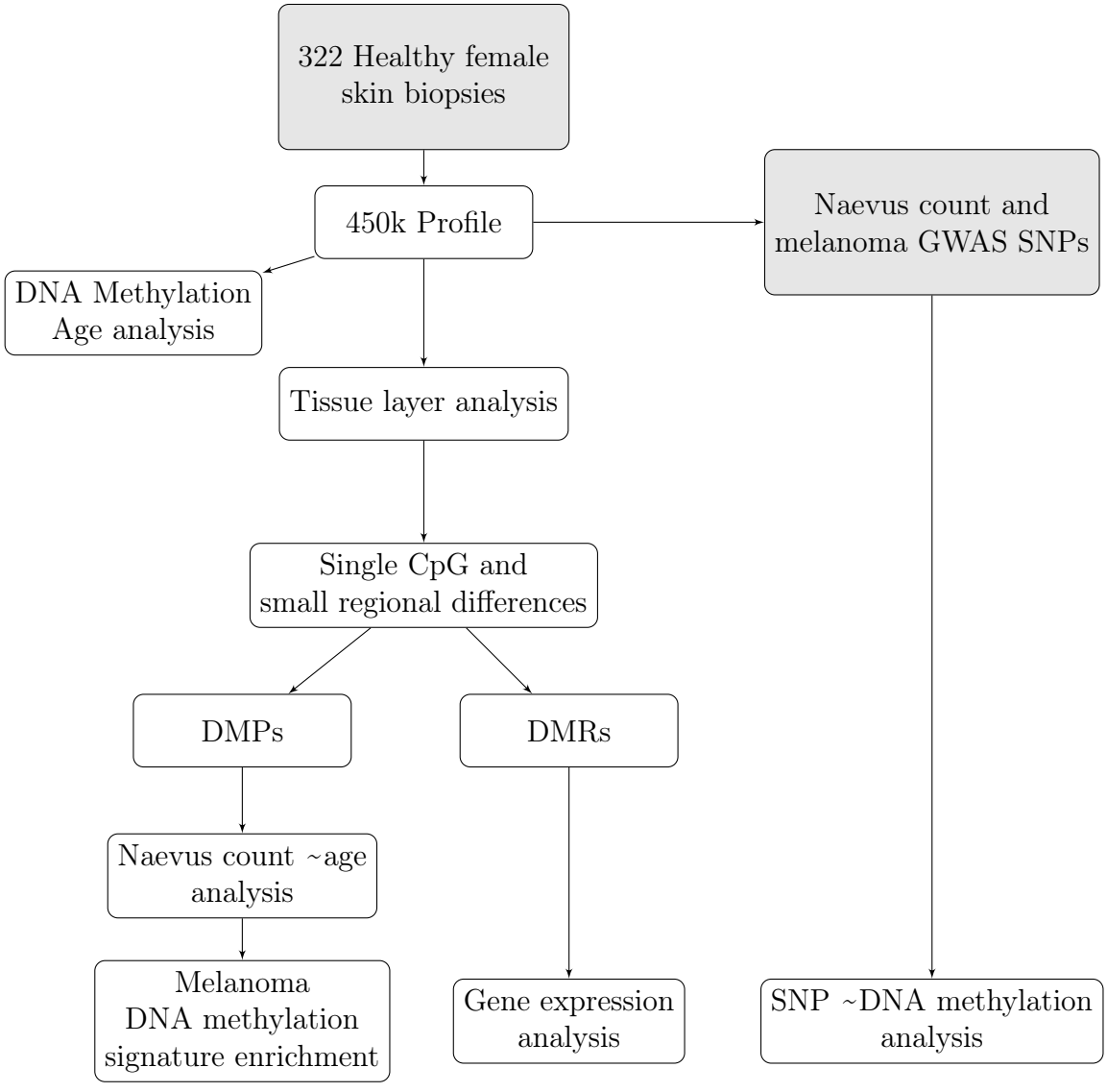



**Figure 5.3: Schematic overview of statistical analyses.** A general overview of the analyses performed in this chapter.

#### Global DNA Methylation Profiles

PCA was performed on the 322 skin samples and 40 dermal and 38 epidermal samples [277] downloaded from the GEO database. Within twin-pair correlations where assessed for all available CpG sites in MZ twin-pairs, DZ twin-pairs, and randomly

paired unrelated individuals using a Spearman's rank test. To test for significant differences in correlation between the groups, a two-sample t-test was performed.

**DNA Methylation Age**

An age prediction model (DNAm age calculator) has been developed by Horvath [32] with high accuracy across a wide range of tissues for use with DNA metylation, either profiled by the 450k or 27k [32]. Here, DNA methylation age was calculated from beta values using this DNAm age calculator webpage (https://dnamage.genetics.ucla.edu/) with the default setting "Normalize Data".

**Naevus Count DMPs**

N-DMPs were identified by testing the association between naevus count and DNA methylation levels at each individual CpG-site. A linear mixed effects model was fitted on the standardised beta values per probe (N(0,1)) and total body naevus count, age, BMI, smoking status, bisulphite conversion efficiency, beadchip, and position on the beadchip were included as fixed effects predictors, and family and zygosity were included as random effects. To assess for significance, an F-test was used to compare this model to a null model without total body naevus count. The significance level was adjusted for multiple testing by use of a FDR of 5% using the "qvalue" package [278] in R, and a suggestive results threshold was set at a FDR of 10%.

**Naevus count DMRs**

N-DMRs were identified using the R package "Bumphunter" [176]. Genomic regions for analysis were set using predefined criteria to require at least three consecutive CpG sites with a maximum gap of 500 bp between the CpG sites. The input DNA methylation levels at the included CpG sites were adjusted for the variables used in the n-DMP analysis. The algorithm determines a $p$ value based here on 1,000 permutations as well as determining FWER adjusted $p$ value. N-DMRs were

identified at a $p$ value $<0.01$ and as epigenome-wide statistically significant at a $p$ value $<0.01$ and at a FWER adjusted $p$ value $<0.05$.

### Age Analysis

Association with age for n-DMPs were tested by using a linear mixed effects model fitted on standardised beta values per probe (N(0,1)) with age, BMI, smoking status, bisulphite conversion efficiency, beadchip, and position on the beadchip as fixed effects, as well as family and zygosity as random effects. To assess for significance, an F-test was used to compare this model to a null model without age.

### Gene Expression Analysis

Gene expression analysis was performed across 248 individuals using gene expression levels available for all genes within 20 kb of each CpG site from the observed n-DMRs. Linear mixed models were fitted on gene expression data with age, BMI, smoking status, batch, and concentration (fixed effects) and family and zygosity (random effects) as well as the linear mixed model fitted for DNA methylation data used for n-DMP analysis described as the null model. This was followed by a Pearson correlation test on the residuals from these models.

### Genetic Variation Analysis

Genetic association analyses of GWAS SNPs with DNA methylation variation in *cis* were performed in a subset of 283 individuals using genome-wide efficient mixed model association (GEMMA), which can account for differing degrees of relatedness. For this, adjusted DNA methylation levels were used accounting for the fixed covariates in the n-DMP analyses. For regional plots, all common SNPs in 100 kb flanking regions of the GWAS SNPs were tested and LocusZoom [279] was used for the regional figures.

### 5.2.7 Genomic Annotation Analysis

All CpG sites included in this chapter were compared with annotations for CpG density (CGI, shores, and shelves) from the UCSC track [177], RefSeq genes (promoter, 5'UTR end, gene body, 3'UTR, intergenic), and functional genomic elements derived from ENCODE including ChromHMM state segmentation, DNase-I hypersensitivity sites, and TFBSs [178, 179].

For each annotation category an enrichment analysis was performed comparing the top 48 ranked n-DMPs to the remainder of CpG sites (n = 415,861). Subsequently, a Fisher's exact test was performed to test significance.

## 5.3 Results

### 5.3.1 The Skin DNA Methylome and Tissue Layer Specificity

The dermis and epidermis have distinct DNA methylomes [277] due to their distinct cell types. This study included skin tissue was obtained from peri-umbilical punch biopsies from 322 healthy female individuals where the adipose and connective tissue layer (hypodermis) was mechanically separated before freezing. To establish the representation of the epidermal or dermal layer in these biopsies, these DNA methylomes were compared to recently published DNA methylomes of separated epidermal and dermal tissue from Vandiver *et al.* [277]. To this end, PCA was performed on the epigenome-wide unadjusted DNA methylation levels of the punch biopsy (n = 322), dermal (n = 40), and epidermal tissue (n = 38). The distinct skin layers were captured by the first two PCs representing 55.6% of the variance as previously reported [277]. The skin tissue biopsy DNA methylomes cluster with the dermal layer DNA methylomes (see Figure 5.4 A), thus the biopsy represents the dermis for the vast majority DNA methylation profiles measured.

The individuals in this sample included complete MZ (n = 25) and DZ (n = 65) twin-pairs as well as unrelated individuals (n = 142). Within twin-pair correlations of MZs and DZs were statistically stronger than correlation of unrelated individual pairs on average ($p = 1.95 \times 10^{-5}$, see Figure 5.4 B) showing the influence of genetic variation. The average MZ within-pair correlation in the skin DNA methylomes ($r_S = 0.986$) is similar to the average MZ twin-pair correlation in peripheral blood described previously in chapter 3, as well as comparable to previously genome-wide estimates [101, 143, 181, 182].

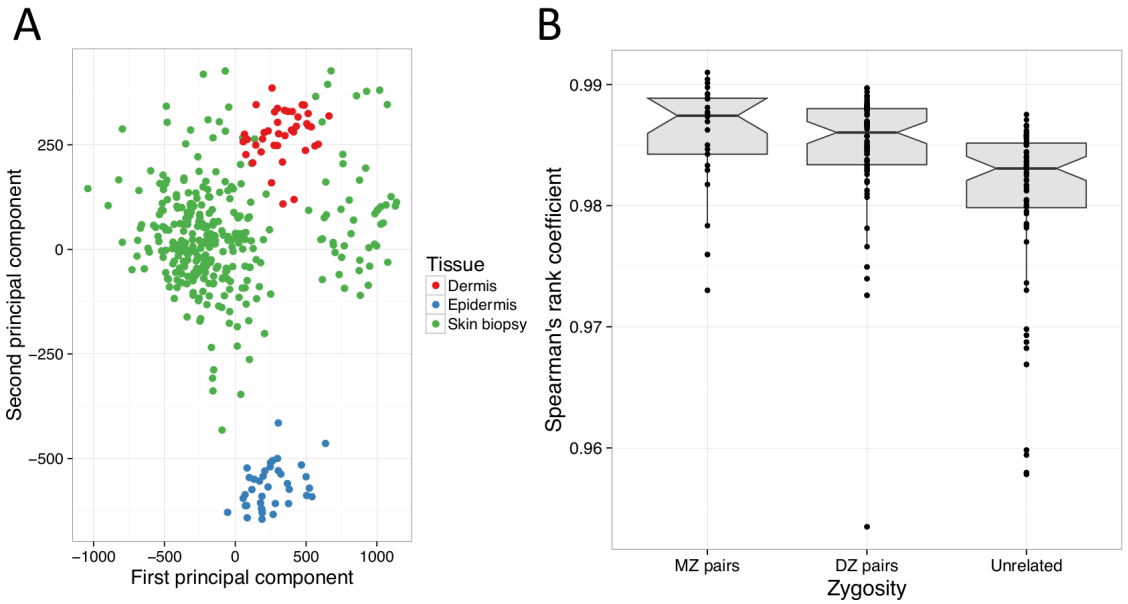



**Figure 5.4: Skin DNA methylome profiles and skin layer specificity. (A)** The first two principal components are coloured for dermal tissue (red), epidermal tissue (blue), and for skin biopsy tissue (green, see legend). **(B)** Pair-wise correlation in DNA methylation profiles shows greater similarity within MZ and DZ twin-pairs compared to pairs of unrelated individuals.

### 5.3.2 DNA Methylation Age Is Not Strongly Correlated With Chronological Age

Recently, Horvath [32] developed an age prediction model with high accuracy across a wide range of tissues for use with either 450k or its predecessor the 27k [32]. Dermal tissue DNA methylomes from 20 individuals were included only in the training set for the prediction model, nevertheless it was identified as one of the tissues where the DNA methylation estimated age showed poor calibration with chronological age with a reported correlation of 0.92 and error of 12 years (defined as the median absolute difference between DNA methylation age and chronological age) [32]. At present, the age calculator has not been applied to large samples of primary skin tissue, and therefore its applicability for this tissue type was assessed here on these data.

In this large dermis sample of 322 individuals, the correlation between chronological age and DNA methylation age was estimated at 0.78 ($p$ = 2.2 x $10^{-16}$), which is lower than the original estimate of 0.92 from Horvath [32], along with a smaller mean error of 6.1 years (see Figure 5.5). The DNA methylation age calculator overestimated age in individuals younger than 50 and consistently underestimated age in older individuals. This trend was also observed in the earlier application in dermal tissue [277]. Furthermore, the slope of 0.57 indicates that the overestimation of DNA methylation age and chronological age increases steadily with age from age 47 and is likely to keep increasing in individuals older than the individuals in this sample (that is, over 80 years of age). Further primary skin tissue studies are needed to confirm this, but these results suggest that the biological ageing is potentially represented differently in DNA methylomes from skin tissue compared to the wide range of other tissues for which the age prediction model works well.

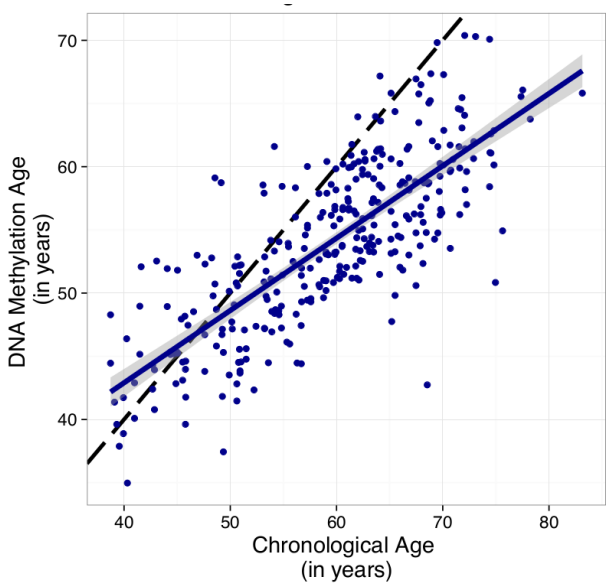



**Figure 5.5: DNA methylation age calculator in skin.** Correlation of chronological age and estimated DNA Methylation age. The black striped line indicates a perfect prediction, the blue line is the true least squares regression fit.

### 5.3.3 Total Body Naevus Count Associated DMPs

The skin DNA methylomes of 322 female individuals were analysed genome-wide for association with total body naevus count at single CpG sites, with the aim of identifying n-DMPs. A linear mixed-effects model was fitted regressing normalised DNA methylation levels on fixed effects (age, BMI, smoking status, chip, order on the chip, and bisulphite conversion efficiency) and random effects (family relatedness and zygosity). This EWAS identified three DMPs associated with total body naevus count at a FDR of 5%. A further 45 n-DMPs were identified at a moderate threshold of FDR 10% (see Figure 5.6 A, Table 5.3, and Supplementary Table S1). These 48 n-DMPs were subsequently assessed for enrichment of CpG density (CGI, shores, and shelves) and ChromHMM state segmentation from ENCODE to remaining number of CpG sites (n = 415,861). Taken together, these 48 n-DMPs were enriched for strong enhancers (state 4) in the epidermal keratinocytes (NHEK) cell line ($p$ = 0.03). Additionally, these n-DMPs were also enriched for CGI shores ($p$ = 0.04) and depleted for "open sea" regions ($p$ = 2.2 x $10^{-3}$).

The three epigenome-wide significant n-DMPs are shown in detail in Figure 5.6 B. These include the most associated n-DMP, cg06244240 ($p$ = 2.1 x $10^{-8}$, FDR 5%), within a CGI shore ~6.5 kb downstream of *METRNL*, a gene involved in glial cell formation that is expressed in healthy skin [280]. The next n-DMP (cg06123942, $p$ = 2.2 x $10^{-7}$) resides within the 5'CGI promoter of *C15orf48*, a gene that has reduced transcript levels in squamous cell carcinomas [281]. Lastly, the n-DMP, cg25384157 ($p$ = 3.1 x $10^{-7}$), is within a CGI shore ~1.5 kb upstream of the TSS of *ARRDC1*, one of the negative regulators of the Notch signalling pathway [282]. This highly conserved pathway is implicated in melanocyte differentiation and is differentially expressed in melanoma [283, 284].

In addition, the n-DMP ranked fourth, cg11297934 ($p$ = 1.2 x $10^{-6}$), is located ~200 bp upstream of the TSS of proto-oncogene *RAF1* (otherwise known

as *CRAF*). This is a member of the *RAF* family, which also includes *BRAF*, in the MAPK/ERK pathway. *BRAF* is mutated in approximately half of all melanomas [285] and is a frequent (driver) mutation in other cancers [286].

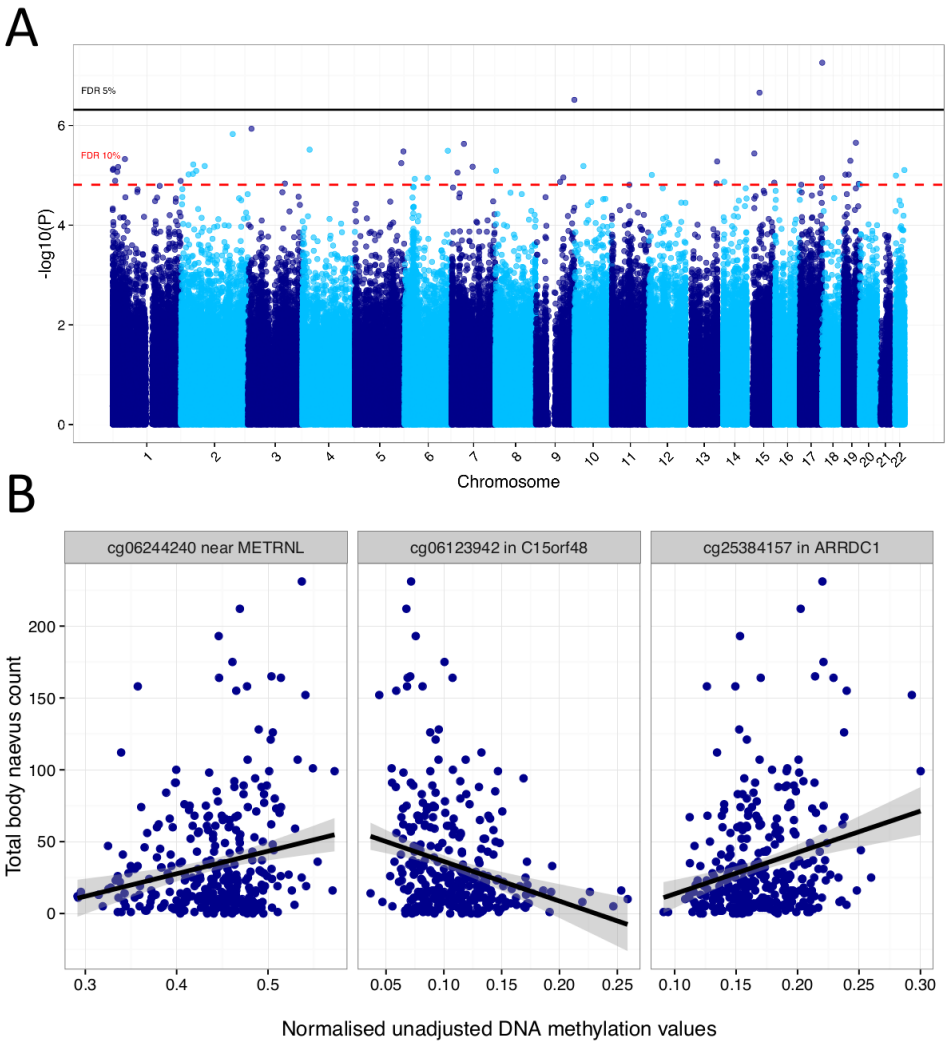



**Figure 5.6: Naevus count EWAS results in 322 female individuals. (A)** Manhattan plot of the EWAS results where each point represents the observed $-\log_{10}$ $p$ value at a single CpG-site. The FDR thresholds of 5% and 10% are depicted as black and red striped vertical lines respectively. **(B)** Panel plot showing the three top-ranked signals for cg06244240 (left), cg06123942 (middle), and cg25384157 (right) using normalised unadjusted beta values per individual. The lines depict a least squares regression fit between DNA methylation and total body naevus count.

**Table 5.3: Most associated DMPs from naevus count EWAS in 322 individuals.**

| Rank | CpG | Position (hg19) | Gene | Location | CpG density | Beta | St. Error | *P* value | FDR |
|---|---|---|---|---|---|---|---|---|---|
| 1 | cg06244240 | chr17:8,1058,948 | - | - | Shore | 0.0052 | $9 \times 10^{-4}$ | $5.5 \times 10^{-8}$ | 5% |
| 2 | cg06123942 | chr15:45,722,795 | *C15orf48* | 5'UTR | Island | -0.0074 | 0.0014 | $2.2 \times 10^{-7}$ | 5% |
| 3 | cg25384157 | chr9:140,499,131 | *ARRDC1* | TSS 1500 | Shore | 0.0063 | 0.0012 | $3.1 \times 10^{-7}$ | 5% |
| 4 | cg11297934 | chr3:12,705,868 | *RAF1* | TSS 200 | Island | -0.0046 | $9 \times 10^{-4}$ | $1.2 \times 10^{-6}$ | 10% |
| 5 | cg14762973 | chr2:187,714,067 | *ZSWIM2* | TSS 200 | Island | 0.0069 | 0.0014 | $1.49 \times 10^{-6}$ | 10% |

**Table 5.4: Most associated DMRs from naevus count EWAS in 322 individuals.**

| Rank | Position (hg19) | Gene | Location | CpG density | Number of CpG sites | DNA methylation | *P* value | Direction CpG sites |
|---|---|---|---|---|---|---|---|---|
| 1 | chr9:140,499,132-140,500,813 | *ARRDC1* | TSS 1500 - Body | Island | 7 | + | $2.5 \times 10^{-5}$ | + + - - - + + |
| 2 | chr10:14,647,154-14,647,530 | *FAM107B* | Body | Shore | 3 | + | $2.5 \times 10^{-4}$ | + + + |
| 3 | chr19:44,285,297-44,285,568 | *KCNN4* | TSS 200 - 1st Exon | - | 3 | + | $2.9 \times 10^{-4}$ | + + + |
| 4 | chr17:8,129,997-8,130,356 | *CTC1* | 3'UTR | Shelf | 3 | - | $6.3 \times 10^{-4}$ | - - - |
| 5 | chr15:26,915,414-26,915,752 | *GABRB3* | Body | Island | 3 | - | $8.3 \times 10^{-4}$ | - - - |

### 5.3.4 Total Body Naevus Count Associated DMRs

Next, differentially methylated regions encompassing multiple CpG sites associated with total body naevus count, n-DMRs, were investigated for the 322 individuals. Here, 48 n-DMRs were identified after 1,000 permutations by BumpHunter ($p$ <0.01) on adjusted DNA methylation levels (see Table 5.4 and Supplementary Table S2). These were then examined regarding their genomic context with a focus on the top five ranked regions.

The most associated n-DMR overlaps a 5' CGI promoter of *ARRDC1* ($p$ = 6.8 x $10^{-5}$, FWER adusted $p$ = 0.05, see Figure 5.7 A). This region encompasses 7 single CpG sites that include the n-DMP cg25384157 that passed FDR 5%, as well as three more CpG sites with a similar direction of effect ($p$ <0.05). Also identified was an n-DMR spanning three CpG sites within a CGI shore in the 5' promoter of an alternative isoform of *FAM107B* ($p$ = 2.5 x $10^{-4}$, FWER adjusted $p$ = 0.25, see Figure 5.7 B). *FAM107B* is expressed in various tissues and is down regulated in multiple cancers [287].

Two more n-DMRs were also identified with the single CpG site EWAS (FDR 10%) and show consistent direction of effect in their neighbouring CpG sites within the n-DMRs in *KCNN4* (5' promoter and TSS, $p$ = 2.9 x $10^{-4}$, see Figure 5.7 C), a potassium channel, and *GABRB3* (first intron, $p$ = 8.3 x $10^{-4}$, see Figure 5.7 E), a GABA receptor. Lastly, a n-DMR was identified in the 3' UTR of *CTC1* and comprised of three CpG sites consistently negatively associated with total body naevus count ($p$ = 6.3 x $10^{-4}$, see Figure 5.7 D). It is also 2 kb upstream of *LINC00324* and lies in a region marked by active promoter states (ChromHMM in multiple ENCODE cell lines including NHEK). *CTC1* is part of the CTS complex that protects telomeres from degradation.

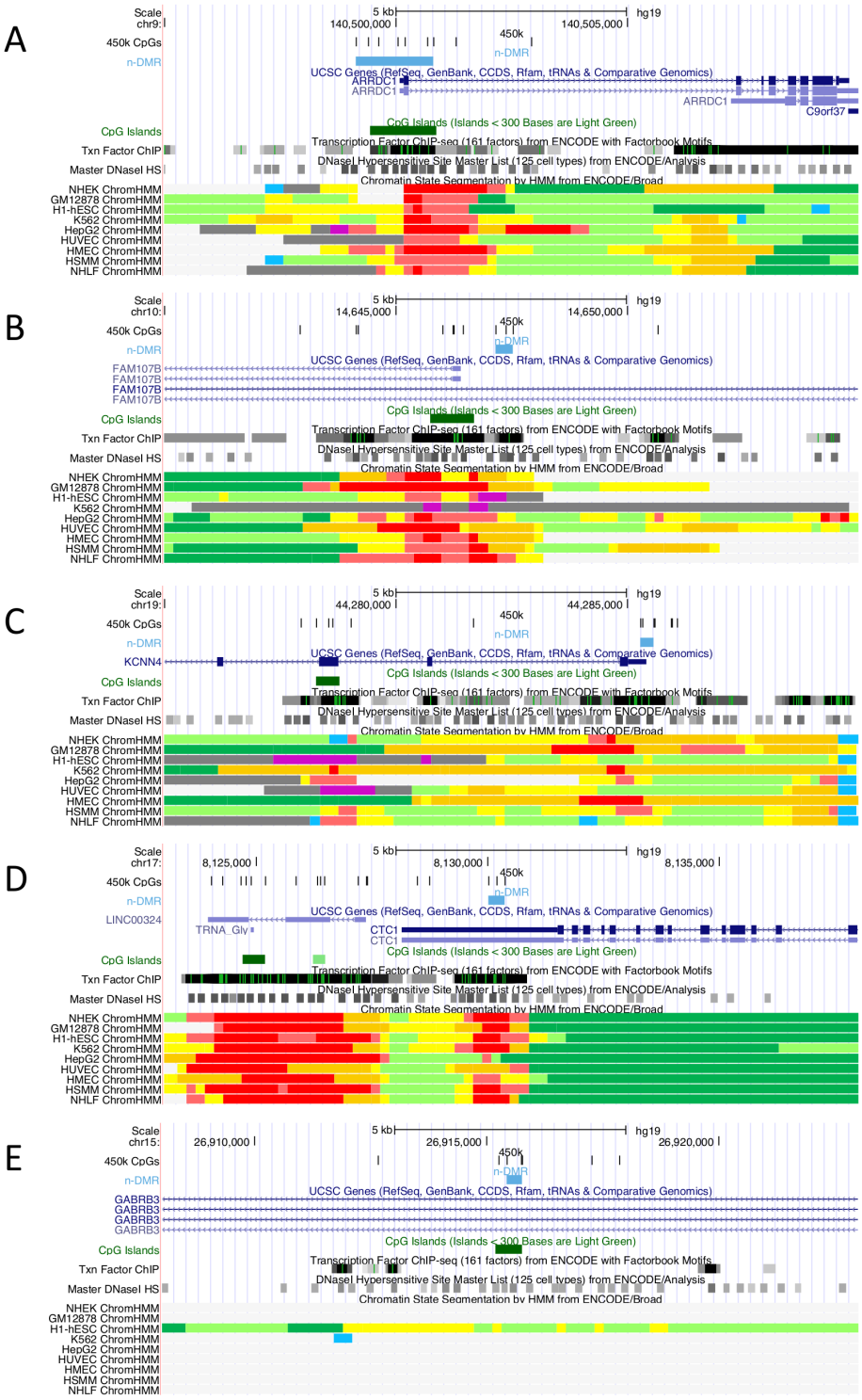



**Figure 5.7: Location of the top five ranked naevus DMRs in the human genome.** Figures obtained from UCSC Genome browser [239], displaying position in the genome (hg19), CpG sites from the 450k, n-DMR (in light blue), RefSeq genes, CGI, transcription factor ChIP data, DNase-I sensitivity sites, and ChromHMM genomic segmentation. **(A)** At *ARRDC1*. **(B)** At *FAM107B*. **(C)** At *KCNN4*. **(D)** At *CTC1*. **(E)** At *GABRB3*.

### 5.3.5 Total Body Naevus Count and Age

Due to the intricate relationship between total body naevus count, melanoma risk, and age, DNA methylation levels were explored at the 48 n-DMPs for a direct association with age, adjusting for the same covariates as previously described. In this sample of 322 individuals, the age range was between 39-82 years old, which is approximately the age when naevi start and continue to involute steadily and linearly. Indeed, a negative correlation of -0.21 ($p$ = 7 x $10^{-5}$, see Figure 5.8 A) was observed between total body naevus count and age.

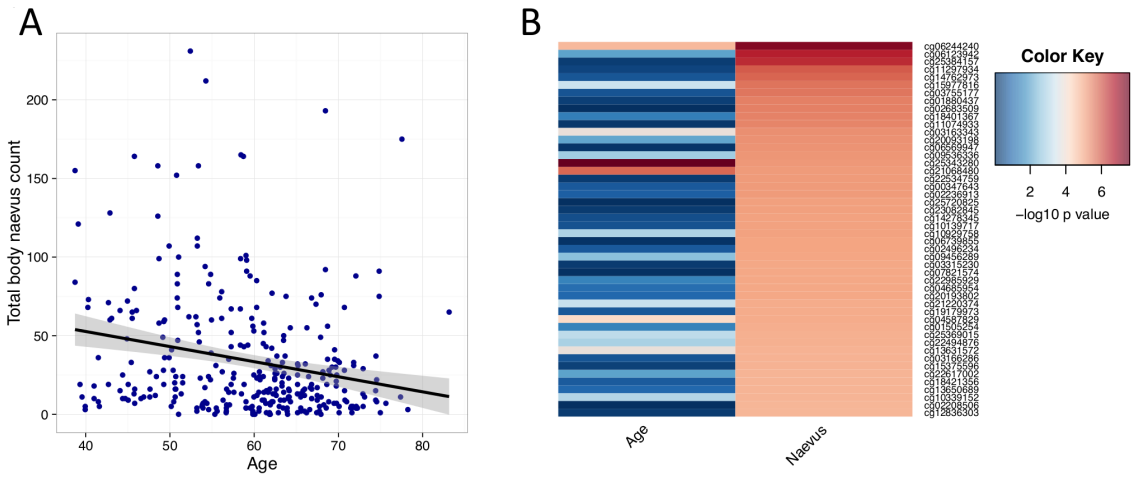



**Figure 5.8: Total body naevus count and age associations.** **(A)** Age at naevus count vs total body naevus count with a least squares regression fit. **(B)** Heatmap of top 48 ranked n-DMPs coloured by – $\log_{10}$ $p$ values of age association and naevus association.

Overall, the 48 n-DMPs had lower $p$ values in the EWAS for naevus count compared than the direct association with age (see figure 5.8 B). Two n-DMPs, cg25343280 and cg21068480, ranked 16 and 17, did show slightly stronger associations with age than total body naevus count. Moreover, no nominally significant associations were detected with age (all $p > 0.05$) at 28 out of the 48 n-DMPs, including *ARRDC1* ($p = 0.57$) and *RAF1* ($p = 0.43$).

As expected, an opposite direction coefficient was identified at these age

associated DMPs considering the negative correlation between age and total body naevus count. Of the top 10 ranked n-DMPs, six did not show any association with age ($p > 0.05$) and the remaining four did not show stronger associations with age. This suggests that the majority of n-DMPs are not directly associated with age.

### 5.3.6 Naevus Count DNA Methylation Signature Is Enriched for Melanoma Associated DNA Methylation Variation

Healthy tissue DNA methylomes can show risk-factor related signatures similar to what has been observed in malignant tissue [127]. Tumour associated DNA methylation changes can also derive from a continuum of maturation states reflecting the normal stages in development [31]. Therefore, the 48 n-DMPs were assessed for enrichment within the DNA methylome changes previously identified in melanoma tumour tissues from two EWAS that compared DNA methylomes of melanoma to either normal skin, melanocytes, or naevi [288, 289].

To date, the most extensive DNA methylome study in melanoma compared the DNA methylomes of melanocytes from three healthy donors and 27 metastatic melanomas, using methylated-CpG island recovery assay sequencing (MIRA-seq). They reported 3,113 hypermethylated regions in melanomas and only 4 hypomethylated regions, as would be expected with a CGI targeted approach. Of the 3,113 hypermethylated regions, 2,039 included at least one CpG site that was also profiled in the 322 individuals analysed in this chapter. These 2,039 regions with increased DNA methylation in metastatic melanomas were further explored for CpG sites that were identified in this chapter as positively associated to total body naevus count ($p$ <0.05). In total, 13.4% (274 regions encompassing 406 CpG sites) were identified as also positively associated to total body naevus count ($p$ <0.05). If only CGIs were considered within these 2,039 regions, enrichment was observed for CpGs identified as positively associated to total body naevus count (Fisher's $p = 6.33 \times 10^{-6}$).

The second EWAS by Koga *et al.* [289] also used a genome-wide based

approach by targeting DNA methylation of 24,103 RefSeq promoters to examine the DNA methylomes of normal skin, naevi, and advanced melanoma. They identified differential promoter methylation at four genes and two of these, *THBS1* and *TNFRSF10D* (hypermethylated in advanced melanoma), were also identified to be positively associated to total body naevus count in this dataset ($p$ <0.05).

### 5.3.7 Total Body Naevus Count DMRs Correlated With Gene Expression

The 48 n-DMRs were examined for associations with gene expression in *cis* in a subset of 248 individuals that also had transcriptomic data available from the same skin punch biopsy. For 27 n-DMRs, gene expression levels were available of 36 genes (64 expression probes) that were within a 20 kb window. Individual CpG sites with a similar direction of effect in the EWAS as the n-DMR were selected to represent the n-DMP. Subsequently, DNA methylation and expression levels both were adjusted for similar biological covariates as well as technical covariates (depending on the data), and compared using Pearson correlation.

At twelve unique n-DMRs, DNA methylation levels at individual CpG sites were correlated with the expression of fourteen unique genes ($p$ <0.05, see Table 5.5). This included the n-DMR at *KCNN4* where one of its three CpGs on the edge in the first exon was positively correlated with DNA methylation (cg15977816, r = 0.19, $p$ = 2.9 x $10^{-3}$). This site is ranked 6th in the EWAS n-DMP results. Additional n-DMRs that have more than one CpG site in the region correlated with one of more gene expression probes of the same gene include n-DMRs at: *MED11*, *C14orf50*, *FAM64A*, *KRT86*, *TTC15*, and *C6orf27*.

**Table 5.5: Significant correlations of naevus count DMRs with expression levels.**

| Rank | Position (hg19) | Gene | Location | DNA methylation | Transcript | Gene | Correlation | *P* value |
|---|---|---|---|---|---|---|---|---|
| 3 | chr19:44,285,297-44,285,568 | *KCNN4* | TSS 200 - 1st Exon | + | ILMN_1709937 | *KCNN4* | 0.19 | 2.9 x $10^{-3}$ |
| 13 | chr2:4,600,947-4,601,053 | - | - | + | ILMN_1692706 | *DCUN1D2* | -0.14 | 0.033 |
| 17 | chr1:151,693,222-151,693,261 | *C1orf230* | TSS 1500 | - | ILMN_1681234 | *TNRC4* | 0.13 | 0.037 |
| 19 | chr8:82,633,130-82,633,568 | *ZFAND1* | TSS 200 - Body | - | ILMN_2087989 | *ZFAND1* | 0.14 | 0.025 |
| 33* | chr17:4,634,804-4,634,804 | *MED11* | TSS 200 - 1st Exon | - | ILMN_1762639 | *MED11* | 0.19 | 2.6 x $10^{-3}$ |
| 35* | chr14:65,016,591-65,016,602 | *C14orf50* | TSS 200 | + | ILMN_2153916 | *HSPA2* | -0.13 | 0.047 |
| 37*** | chr17:6,347,533-6,347,533 | *FAM64A* | TSS 1500 | - | ILMN_2415292 | *AIPL1* | -0.18 | 5.5 x $10^{-3}$ |
| 39 | chr16:28,565,199-28,565,206 | *CCDC101* | TSS 200 | - | ILMN_1810560 | *P8* | 0.13 | 0.035 |
| 42 | chr22:39,712,694-39,712,730 | *RPL3*;*SNORD83A* | Body; TSS 1500 | + | ILMN_1653927 | *SNORD83A* | 0.13 | 0.034 |
| 46* | chr12:52,695,412-52,695,515 | *KRT86* | TSS 200 | - | ILMN_1801442 | *KRT81* | -0.19 | 3.2 x $10^{-3}$ |
| 47* | chr2:3,471,345-3,471,345 | *TTC15* | Body | + | ILMN_1693317 | *TTC15* | -0.14 | 0.024 |
| 48* | chr6:31,743,928-31,743,952 | *C6orf27* | Body | + | ILMN_1696601 | *VARS* | -0.18 | 5.5 x $10^{-3}$ |

* Multiple CpG sites associated with multiple transcript probes of same gene, **one CpG site associated with multiple transcript probes of same gene, or *** multiple CpG sites associated with one transcript probe; the strongest association is shown for the n-DMR.

### 5.3.8 Impact of GWAS SNPs for Naevus Count or Melanoma Risk on DNA Methylation in *cis*

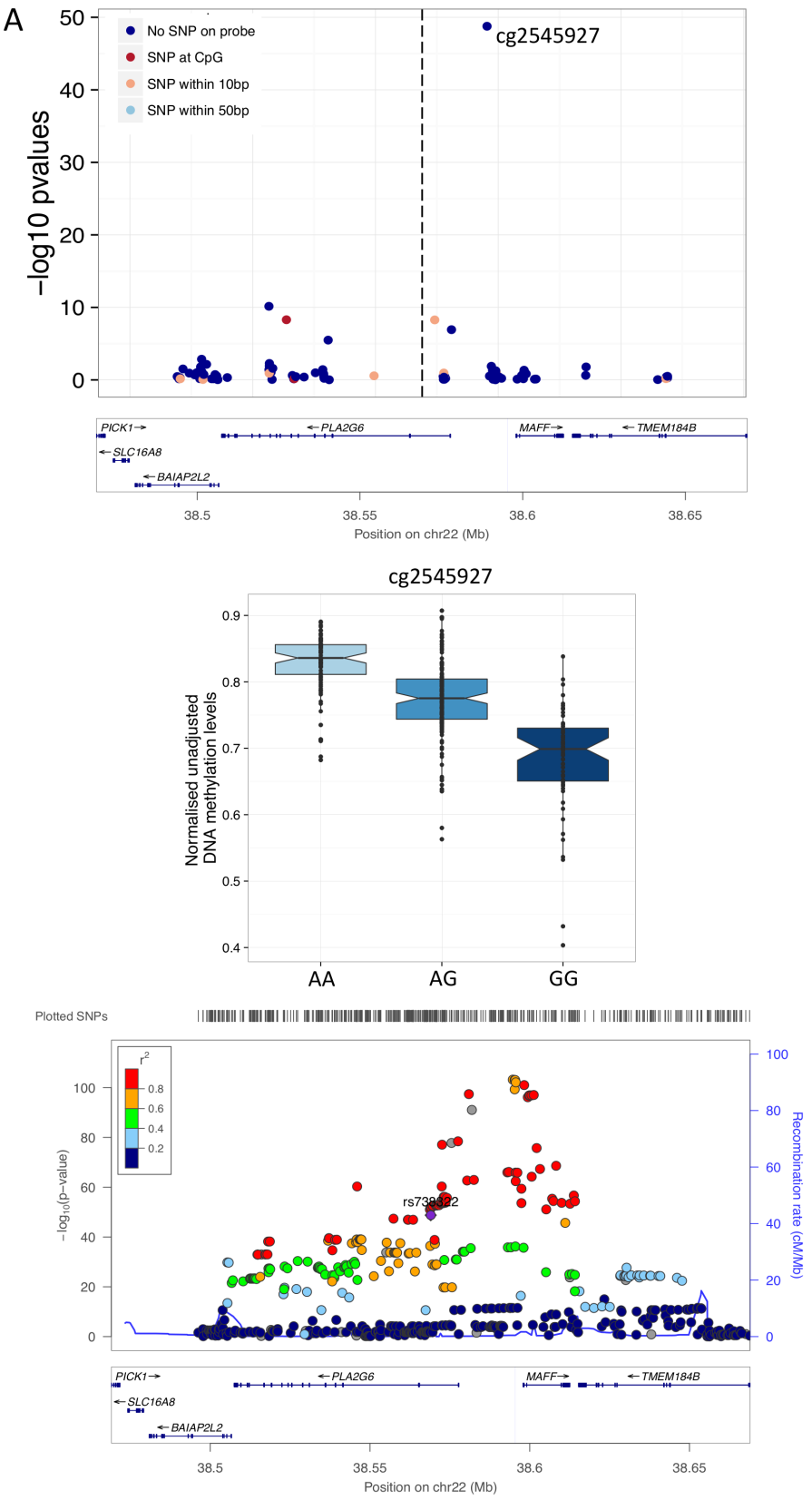
SNPs identified via GWAS for association with total body naevus count or melanoma risk were also investigated for their impact on DNA methylation levels in *cis* for a subset of 283 individuals. Four out of five SNPs associated with total body cutaneous naevus count [272, 273] and all 23 unique SNPs associated with melanoma risk from the GWAS catalogue [290] had at least one CpG site within a 100 kb window of the SNP. To account for differing degrees of relatedness, mixed linear models were performed to test for genetic association between the SNP variant and DNA methylation levels. The multiple testing threshold was set at $p < 1.0 \times 10^{-5}$ which is comparable to an EWAS performed using the 450k where all CpG sites were assessed with common genetic variants (MAF >1%) within a 100 kb window ($p < 8.6 \times 10^{-4}$) by Grundberg *et al.* [143].

Altogether thirteen GWAS SNPs were associated with DNA methylation levels at CpG sites in *cis* in skin tissue ($p < 1 \times 10^{-5}$, see Table 5.6). These included three out of the four SNPs from naevus count GWAS results: rs2284063 in the second intron of *PLA2G6* (cg25457927, $p = 3.5 \times 10^{-38}$), rs738322 in the first intron of *PLA2G6* (cg25457927, $p = 1.7 \times 10^{-49}$, see Figure 5.9 A), and rs3768080 in the tenth intron of *NID1* (cg18765906, $p = 6.4 \times 10^{-14}$, see Figure 5.9 B). The remaining ten SNPs associated with DNA methylation were from melanoma risk GWAS results and these variants located in or near the genes: *MC1R* (2 SNPs), *MX2* (see Figure 5.9 C), *TERT/ CLPTM1L* (see Figure 5.9 D), *PLA2G6*, *CASP8*, *ACTRT3*, *ASIP*, *CDC91L1*, and *ARNT/ SETDB1/LASS2ANXA9/ MCL1/ CTSK*. Of the most associated CpG sites per thirteen SNPs, six CpG sites were in active promoters or enhancers in NHEK cell line [178]. None of the associated CpG sites were epigenome-wide significant in the EWAS of single CpG sites or of small regions. However, nine were nominally significant ($p < 0.05$) including *PLA2G6* which is associated with both naevus count and melanoma risk by GWAS [272, 274].

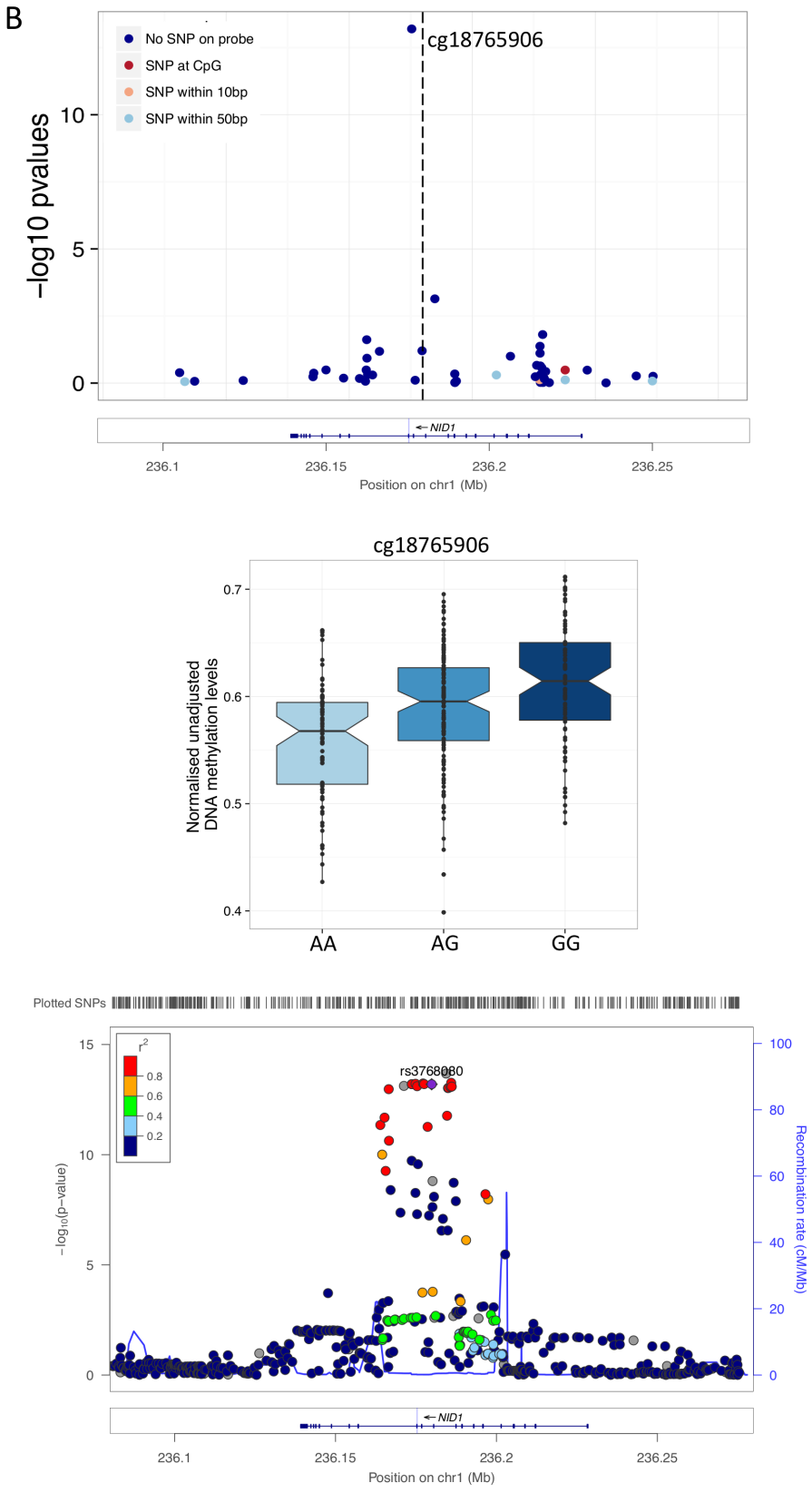
Subsequently, all common SNPs (MAF >1%) in 100 kb flanking regions of these thirteen GWAS SNPs were investigated for influence on the same CpG sites. The *PLA2G6* GWAS SNPs (rs738322) is part of a large linkage disequilibrium (LD) block that spans over the promoter sites of both *PLA2G6* and *MAFF*. In this case the significant CpG site (cg25457927), ~2 kb upstream of *MAFF*, in fact showed a much stronger association with other SNPs within this block ($p < 1.0 \times 10^{-100}$) (see Figure 5.9).

For eight of these thirteen SNPs, the Genotype-Tissue Expression Project (GTEx) also identified expression quantitative trait loci (eQTLs) in skin tissue (sun or not sun exposed) and/or in transformed fibroblasts. In total, these eQTLs were associated with the expression 16 genes, seven at which DNA methylation variation was also associated at the same gene with the same SNP; *CASP8* (rs1301693), *MAFF*, *PLA2G6*, *TMEM184B*, and *BAIAP2L2* (rs2284063, rs738322, and rs6001027), *SPATA33* (rs258322), *MX2* (rs45430), and *CDK10* (rs4785763).

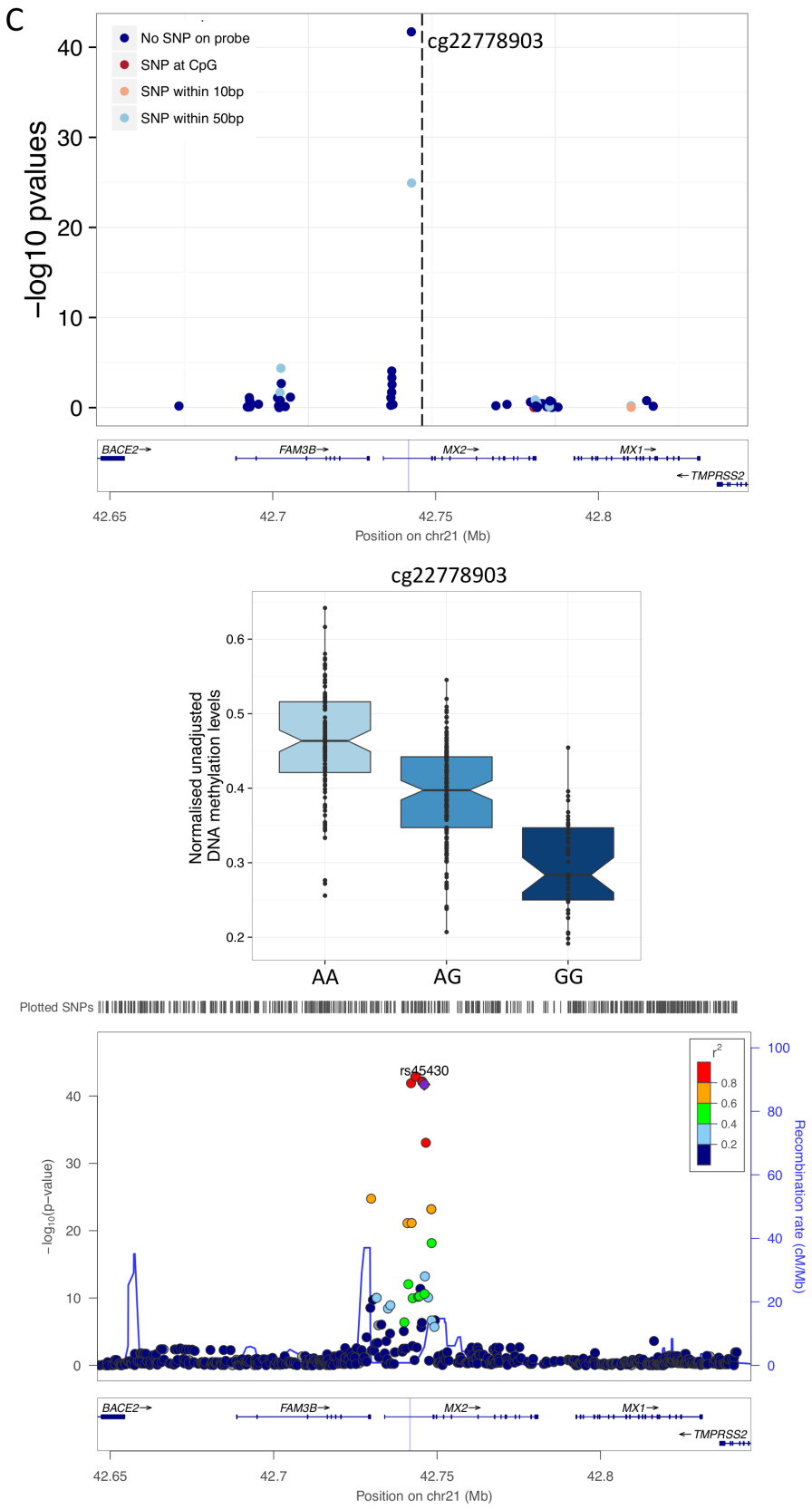
**Figure 5.9: Genetic locus regions for GWAS SNPs and DNA methylation variation in *cis* shown on the following four pages**. Per page, a panel plot comprises three figures each. **(Top)** Regional plot of 100 kb flanking regions around the GWAS SNP of interest indicated by a striped black line and each point is a CpG with its -$\log_{10}$ $p$ value on the y-axis. Each point is coloured according to occurrence of genetic variants on the probe sequence shown in the legend. **(Middle)** The middle boxplot shows the strongest associated CpG site for each genotype using normalised unadjusted DNA methylation levels. **(Lower)** Plot of same region generated by LocusZoom [279] showing all SNPs in the region against the strongest CpG site.



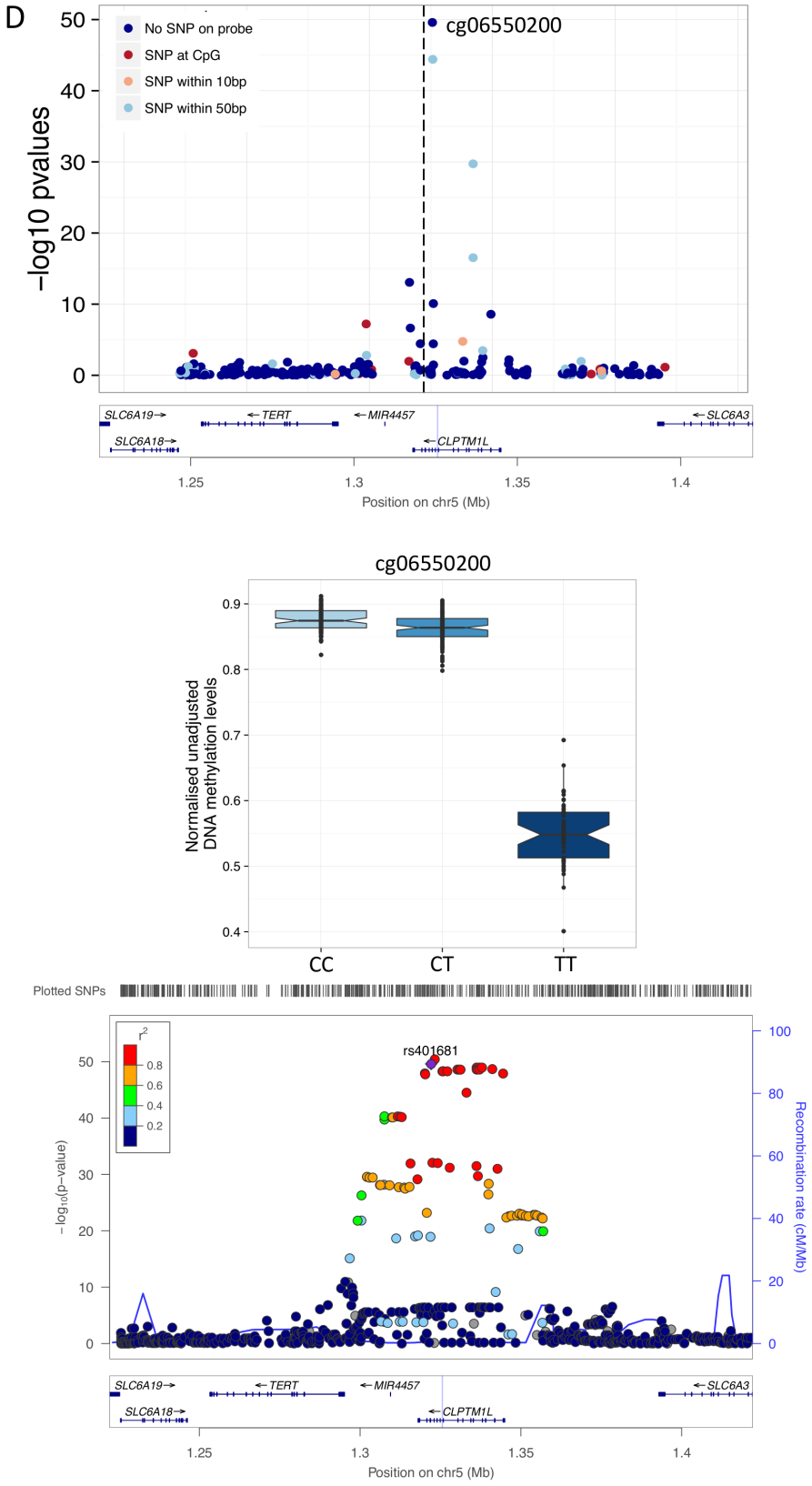

**5.9(A)** rs738322 identified for cutaneous naevi.




**5.9(B)** rs3768080 identified for cutaneous naevi.




**5.9(C)** rs45430 identified for melanoma risk.




**5.9(D)** rs401681 identified for melanoma risk.

**Table 5.6: Strongest CpG association per GWAS SNP.**

**Begin of multi-page Table**

| Trait | rs | Position (hg19) | Reported Genes | CpG | Beta | *P* value | Gene | Location | CpG density | Distance to SNP (kb) |
|---|---|---|---|---|---|---|---|---|---|---|
| Melanoma | rs401681 | chr5:1,322,087 | *TERT*; *CLPTM1L* | cg06550200 | 0.783 | $2.6 \times 10^{-50}$ | *CLPTM1L* | Body | - | -3.5 |
| Cutaneous naevi | rs738322 | chr22:38,569,006 | *PLA2G6* | cg25457927 | 0.851 | $1.7 \times 10^{-49}$ | - | - | Shelf | -26.4 |
| Melanoma | rs45430 | chr21:42,746,081 | *MX2* | cg22778903 | -0.711 | $1.9 \times 10^{-42}$ | *MX2* | 5'UTR | - | 4.4 |
| Melanoma | rs6001027 | chr22:38,545,619 | *PLA2G6* | cg25457927 | 0.843 | $1.3 \times 10^{-38}$ | - | - | Shelf | -49.8 |
| Cutaneous naevi; Melanoma | rs2284063 | chr22:38,544,298 | *PLA2G6* | cg25457927 | 0.834 | $3.5 \times 10^{-38}$ | - | - | Shelf | -51.1 |
| Melanoma | rs7412746 | chr1:150,860,471 | *ARNT*; *SETDB1*; *LASS2*; *ANXA9*; *MCL1*; *CTSK* | cg15448220 | 0.662 | $1.4 \times 10^{-32}$ | *SETDB1* | TSS1500 | Shore | -37.4 |
| Melanoma | rs258322 | chr16:89,755,903 | *MC1R* | cg05714116 | -0.728 | $5.5 \times 10^{-14}$ | *CDK10* | TSS1500 | Shore | 3.3 |
| Cutaneous naevi | rs3768080 | chr1:236,179,869 | *NID1* | cg18765906 | -0.374 | $6.4 \times 10^{-14}$ | *NID1* | Body | - | 4.5 |
| Melanoma | rs4785763 | chr16:90,066,936 | *MC1R* | cg08547343 | -0.346 | $1.0 \times 10^{-9}$ | *CENPBD1*; *AFG3L1* | 5'UTR; TSS200 | Island | 28.1 |
| Melanoma | rs910873 | chr20:33,171,772 | *CDC91L1* | cg01901788 | -0.441 | $1.8 \times 10^{-6}$ | *MAP1LC3A* | TSS1500 | Shore | 25.9 |
| Melanoma | rs13097028 | chr3:169,464,942 | *ACTRT3* | cg27020690 | -0.289 | $2.0 \times 10^{-6}$ | - | - | Island | -17.4 |
| Melanoma | rs13016963 | chr2:202,162,811 | *CASP8* | cg24599065 | -0.23 | $2.1 \times 10^{-6}$ | *CASP10* | 3'UTR | - | 69 |

**Continuation of Table on next page**

**Continuation of Table 5.6**

| Trait | rs | Position (hg19) | Reported Genes | CpG | Beta | *P* value | Gene | Location | CpG density | Distance to SNP (kb) |
|---|---|---|---|---|---|---|---|---|---|---|
| Melanoma | rs228437 | chr6:134,898,456 | *ASIP* | cg24504307 | -0.355 | $9.2 \times 10^{-6}$ | - | - | - | -64.7 |
| Melanoma | rs3219090 | chr1:226,564,691 | *PARP1* | cg18764804 | 0.24 | $6.2 \times 10^{-4}$ | *PARP1* | TSS1500 | Shore | -32 |
| Melanoma | rs1031925 | chr3:51,379,274 | *DOCK3* | cg09456445 | -0.315 | $8.5 \times 10^{-4}$ | *DOCK3* | 3'UTR | Shore | -41.1 |
| Melanoma | rs1722784 | chr1:150,961,869 | *ANXA9* | cg07479786 | 0.261 | $8.9 \times 10^{-4}$ | *ANXA9* | 3'UTR | - | -6 |
| Melanoma | rs16953002 | chr16:54114824 | *FTO* | cg01083598 | -0.169 | $7.2 \times 10^{-3}$ | - | - | Island | -41.2 |
| Melanoma | rs35390 | chr5:33955326 | *SLC45A2* | cg01990593 | 0.689 | $8.8 \times 10^{-3}$ | *ADAMTS12* | Body | Shore | 65 |
| Melanoma | rs1801516 | chr11:108175462 | *ATM* | cg08954307 | -0.249 | $9.8 \times 10^{-3}$ | *ATM* | Body | - | -59.3 |
| Cutaneous naevi | rs4636294 | chr9:21747803 | *MTAP* | cg03724238 | -0.128 | 0.013 | - | - | Island | 51.1 |
| Melanoma | rs4698934 | chr4:106139387 | *TET2* | cg08530497 | -0.195 | 0.037 | *TET2* | Body | - | -15.9 |
| Melanoma | rs1847134 | chr11:89005253 | *TYR* | cg25941151 | -0.153 | 0.041 | *TYR* | TSS200 | - | 94.3 |
| Melanoma | rs7023329 | chr9:21816528 | *CDKN2A* | cg14548963 | 0.118 | 0.068 | *MTAP* | Body | - | 3.6 |
| Melanoma | rs1393350 | chr11:89011046 | *TYR* | cg03508346 | -0.118 | 0.142 | *NOX4* | 3'UTR | - | -48.8 |
| Melanoma | rs17119461 | chr10:107516352 | *NR* | cg18758405 | -0.29 | 0.333 | - | - | - | 65 |
| Melanoma | rs1889497 | chr6:65432283 | *EYS* | cg11999886 | -0.026 | 0.76 | *EYS* | Body | - | 90.8 |

**End of Table**

## 5.4 Discussion

In this chapter, skin DNA methylomes of 322 healthy females were investigated in relation to total number of naevi across the body, the strongest risk factor for melanoma. This is the first study to explore DNA methylation in healthy skin tissue for naevus count. Moreover, the study was performed using the largest available skin tissue dataset to date. It provides an extra layer to our current understanding of the genomic biology behind the number of naevi in individuals on top of that previously obtained via GWAS. The healthy skin biopsies were first shown to represent the dermal layer via their epigenetic signature. Then DNA methylation variation, n-DMPs and n-DMRs, were identified at genes novel to naevi biology as well as known genes in naevi formation or melanoma. The top most associated results were enriched for strong enhancers in NHEK cell lines of ENCODE and for CGI shores, known to be more dynamic and functional regions in both cancer [52] and stem cell reprogramming [54]. Furthermore, these sites were not directly associated with age. Approximately half of the n-DMRs were correlated with gene expression in *cis* within the same biopsy. Finally, DNA methylation variation in *cis* associated with known GWAS SNPs for both naevi number and melanoma risk.

Many differentially methylated genomic loci were identified with total body naevus count and a substantial subset were highly relevant in melanocyte biology, melanoma, and cancer in general. Three are highlighted in this section. One of these, the n-DMR in *CTC1*, involves an already established link between longer telomeres and higher total body naevus count as well as melanoma risk [269]. This gene plays a role in telomere maintenance and has been associated with telomere length via GWAS [291]. Moreover, a genetic score from seven SNPs associated with telomere length have been shown to be robustly associated with melanoma risk [292]. This suggests that healthy skin may also have longer telomeres in individuals

with high naevus counts and supports previous work on the importance of this terminal chromosomal region in melanoma biology [293, 294]. This may also link the observations of lesser photo-ageing of the skin in individuals diagnosed with melanoma (excluding head and neck).

The n-DMP in *RAF1* draws attention to the well known and important genetic pathway in melanoma, the MAP/ ERK pathway, which includes the proto-oncogene *BRAF*. *BRAF* is mutated and subsequently activated in approximately half of all melanomas and its mutational presence or absence influences major therapeutic decisions [285]. The oncogene *RAF1* on the other hand, is rarely mutated in human cancers [295]. It is however, a therapeutic target and leads to apoptosis in *BRAF* negative melanomas. Here, hypomethylation at the TSS of *RAF1* is found with an increase in the number of naevi, and could be an indication of an altered pathway in skin that has an increased disposition to the formation of naevi. Additionally, it could be of use in addition to using mutational status of *BRAF* alone in clinical assessment of advanced melanoma [112, 190, 296].

Increased DNA methylation levels were also observed at the promoter of *ARRDC1*, both epigenome-wide significant at a single CpG and regional level. *ARRDC1* is part of a highly conserved pathway in cell signaling, NOTCH, that is a key player in embryogenesis for cell-fate determination for many organ systems as well as in tissue maintenance. Low expression levels have been identified of *Notch1*, one player in the NOTCH pathway, in melanocytes and naevi and higher expression levels have been identified and associated with melanoma pathogenesis [283, 284]. The role of *ARRDC1* is pivotal in Itch E3 ubiquitin ligase mediated NOTCH receptor degradation [282]. Increased DNA methylation at the promoter of *ARRDC1* is observed in this study makes an intriguing connection with NOTCH pathway.

Enrichment for DNA methylation changes previously identified in

melanomas [288, 289] was observed in this study's naevi-associated DNA methylome signature in normal skin tissue. This may represent a priming or predisposition in skin with higher numbers of naevi to melanoma as risk factor associated DNA methylation changes, such as BMI and smoking, have already been observed in healthy colon as well as colon tumours [127].

Approximately half of the 26 unique SNPs from GWAS for the number of naevi or melanoma risk (from the GWAS catalogue [272, 273, 290]) were associated with DNA methylation variation in *cis* within 100 kb. These included three out of four SNPs previously associated with naevi numbers that are reported at two genes, *PLA2G6* and *NID1*. *PLA2G6* is robustly linked to both naevi numbers and melanoma risk by three unique SNPs. For all three SNPs, eQTLs have also been reported either in skin and/or transformed fibroblasts, impacting the expression of not only *PLA2G6*, but also neighbouring *MAFF*, *TMEM184B*, and *BAIAP2L2*. All of these genes also showed differential methylation associated with these SNPs in this study, highlighting the usefulness of including DNA methylation variation with genetic variants identified by GWAS. Moreover, DNA methylation variation was also associated with SNPs in novel genes without previously reported eQTLs.

One of the limitations of this study is the absence of replication or validation of these results. These steps are paramount in arriving at robust findings and EWAS is no exception. Unfortunately, a skin biopsy is an invasive technique and very few healthy tissues have been collected. To date, none exist that have been profiled for genome-wide DNA methylation as well as detailed naevus counts performed on these individuals. In this chapter all possible individuals were used to maximise the power to detect true DNA methylation modifications, rather than divide the dataset to attempt to provide internal replication. However, new biopsies are being collected in the TwinsUK as part of a follow up of the same individuals that could provide new insights by performing for example longitudinal analysis for these

results. Future studies could aim to profile the DNA methylomes of these skin biopsies and perform longitudinal analyses for these individuals or provide replication of independent samples.

This study is the largest healthy skin DNA methylome study to date [277, 297, 298], and also benefits from the extensive data for naevus counts collected by trained dermatology nurses of healthy individuals. It also has the advantage of including only women, as they have a different pattern of naevi distribution across the body [264] as well as known (autosomal) DNA methylation differences compared to men [299]. In contrast to blood-based EWASs, this EWAS investigates a tissue directly associated with the phenotype of interest and can offer an in-depth view of biological pathways implicated with the phenotype of naevus count.

## 5.5 Conclusion

This novel study has investigated naevi number in 322 healthy females using the DNA methylome, transcriptome, and genetic variants from GWAS for the same individuals. Novel genes and pathways were identified as well as genes known to be involved in melanocyte biology or melanoma progression. This thereby adds additional information to the genetic basis and biological processes underlying the number of naevi and melanoma pathogenesis. These findings may open new pathways to explore, to understand not only why some individuals have higher numbers of naevi, but how this phenotype contributes to melanoma risk.

# Chapter 6

# Discussion

Over the last decade, epigenomic analysis has gained momentum with the implementation of array and second generation sequencing techniques. Epigenomic maps across human tissues and cell lines of DNA methylation, histone modifications, and DNase-I sensitive sites have been generated by large projects now coordinated in the International Human Epigenome Consortium (IHEC), such as ENCODE [300], Roadmap Epigenomics Program [301], and BLUEPRINT [302]. These reference datasets have identified key cellular states associated with healthy tissues, potential mechanisms relevant to diseases, and provide an encyclopedia of these elements across the genome. EWAS have contributed an additional wealth of data regarding variability associated with diseases, environmental factors, and ageing.

This thesis adds to the current state of knowledge by investigating the DNA methylome for systemic changes associated with cancer pathogenesis that are independent of genetic variation, and specific changes in skin associated with cancer risk factors. This was achieved via two main research questions:

1. Analysis of peripheral blood DNA methylomes in cancer discordant MZ twin-pairs for potential biomarkers of disease or disease risk.

2. Analysis of skin DNA methylomes to investigate the skin's predisposition to the number of naevi across the body, the strongest risk factor for melanoma.

## 6.1 Peripheral Blood DNA Methylome Changes in Cancer Discordant MZ Twin-pairs

DNA methylomes of peripheral blood or specific blood cell types have already been investigated and associated with a range of cancer risk factors [33, 66, 68–70, 164–166] as well as with cancer development at primary locations including breast [134, 137, 217–222], colon [138], bladder [139], and ovary [140]. However, the vast majority of these studies have used a population-based design. Whilst this does not diminish the value of identified biomarkers, causal genetic variation underlying these differences in DNA methylation cannot be excluded. Here, the aim was to determine the presence of "pure" DNA methylome changes, *i.e* excluding obligatory or facilitated DNA methylation associations [60] by using discordant MZ twin-designs.

In the first research chapter of this thesis, pan-cancer CpG site-specific changes in peripheral blood DNA methylomes were identified by using 41 cancer discordant MZ twin-pairs. This is the first study to date to investigate peripheral blood DNA methylome changes not specific to only one type of cancer, but focused on identifying pan-cancer biomarkers. The results included one epigenome-wide significant DMP and a further three DMPs that passed a suggestive significance threshold. These candidate regions are of interest for further research into their full potential as the identified changes were present years prior to cancer diagnosis.

In the second research chapter, breast cancer specific DNA methylation regions were identified through two epigenome-wide DNA methylation profiling techniques: 450k for 28 breast cancer discordant MZ twin-pairs and MeDIP-seq for 26 twin-pairs. The sample size here is approximately double of that included in the only previous breast cancer discordant MZ twin-pair study [137]. This work sets itself apart by investigating only DNA methylation changes obtained from blood samples preceding diagnosis. Additionally by using the two technologies to assay the breast cancer blood methylome, genome coverage was significantly improved. The

results included three epigenome-wide significant novel bc-DMRs, two suggestive bc-DMRs, and four suggestive bc-DMPs. These candidate regions are of particular interest because they are present in peripheral blood preceding diagnosis. Their potential requires further research in future prognostic and diagnostic studies.

The most associated peripheral blood DNA methylome signatures for both the pan-cancer and early breast cancer studies did not overlap. This could be a result of the limited number of twin-pairs used in each analysis. Each study presented in this thesis is the largest available to date, however to reach 80% power to identify genome-wide significant DNA methylation changes, more discordant MZ twin-pairs are needed. This highlights the necessity for EWAS to replicate results in other cohorts. The regions and CpG sites identified in these chapters should undergo further analysis to explore their potential application as biomarkers.

## 6.2 Skin DNA Methylome Changes Association With Naevus Count

The second part of this thesis focused on identifying changes in the DNA methylome associated with the skin's disposition to greater number of naevi. This work built on previous observations that healthy tissue DNA methylomes can show risk-factor related signatures that are similar to what has been observed in malignant tissue [127]. Therefore, these DNA methylation signatures can potentially indicate altered mechanisms of disease development. The healthy human skin DNA methylome had not been previously investigated for a potential risk factor signature of increased total body naevus count.

Here, this novel EWAS for total body naevus count identified 48 n-DMPs and 48 n-DMRs in 322 healthy female individuals. Approximately half of the n-DMRs were also correlated with gene expression in *cis* from the same biopsy. Fur-

thermore, DNA methylation variation in *cis* was identified for the majority of GWAS SNPs for both naevus count and melanoma risk.

This is not only the first study for total body naevus count but it also uses the largest available healthy skin DNA methylome dataset. The skin DNA methylome signature associated with naevus count comprises novel genes and pathways as well as genes known to be involved in melanocyte biology or melanoma progression. This helps support both the validity of the design of the study as well as contributing new information regarding risk factor DNA methylation status at these genes. This identified naevus count signature can help to understand how this phenotype might arise and contribute biologically to melanoma risk.

## 6.3 Thesis Strengths

The strength of both chapters on biomarkers in peripheral blood lies partly in the use of a discordant MZ twin-pair design. This design can detect changes independent of host genetic variation and could therefore find CpG sites more vulnerable to environmental factors compared to population-based designs that are influenced by underlying genetic structure. It benefits from the advantage of the TwinsUK cohort that has blood samples stored that were obtained prior to diagnosis. These types of samples are generally only available from cohorts that routinely collect or have stored blood samples of healthy individuals. TwinsUK also has the advantage that it is linked to the ONS for official cancer diagnosis and cause of death (if deceased). Questionnaire data on cancer diagnosis was available, however, it did not correspond with the official records in approximately a quarter of cases. This discrepancy shows a trend towards misreporting in questionnaires that is similar to what is observed for chronic diseases, where misreporting was found to be considerable and differed by respondents level of education [303].

One of the major strengths for the integrative skin DNA methylome study

is the use of the tissue of interest to melanoma and naevus count. This allowed interpretation beyond surrogate changes that might be observed in peripheral blood, and offers a more in-depth biological view of total body naevus count. TwinsUK is one of the few cohorts in the world with total body naevus count data that is collected by dermatology research nurses via an established and repeatable protocol. It also profited from the available expression data from the same biopsy that the DNA methylation was assayed in and that a large number of individuals were genotyped allowing integration of these with data.

Finally, the TwinsUK cohort is a deeply phenotyped cohort. As such data for was available measured over multiple time points in important factors such as BMI, smoking status, and drinking habits.

## 6.4 Thesis Limitations

Ideally, greater numbers of cancer discordant twin-pairs for both the discovery analysis and replication analysis would increase the power of detecting true differences in DNA methylation. Unfortunately, blood samples obtained before diagnosis of MZ twin-pairs are rare worldwide. The future potential of combined consortia could enable pooling of more samples across cohorts for more powerful analyses. It is now well acknowledged that an important factor to consider in EWAS is cellular heterogeneity, particularly for peripheral blood samples, in order to reduce results that are in fact due to cell composition. Cellular heterogeneity in the peripheral blood samples was estimated and adjusted for in the 450k analyses in this thesis, but it was not corrected for in the main reported breast cancer results assessed by MeDIP-seq as blood cell counts from the same time point were not available for the majority of samples. Detection of cell type composition changes or higher levels of rare subtypes [79] does not diminish the potential value as a biomarker if this composition is consistently seen in individuals and if it is cost-effective in the clinic.

Although further analyses in specific cell types should also be performed for more detailed biological interpretation of the results.

Regarding the MeDIP-seq dataset, an increased read coverage would have increased the power of the study by improving the robustness of the data. This dataset could also have benefited by an addition of methylation-sensitive restriction enzyme sequencing (MRE-seq) on the same samples that would enable an approximate prediction of DNA methylation at regions more comparable to 450k and WGBS [304, 305].

The skin DNA methylome study could be improved by replication and validation of the results. Unfortunately, no healthy skin dataset exists to date that has been profiled for genome-wide DNA methylation and that also possesses the phenotype information of detailed naevus counts. In this study, all possible individuals were used to maximise the power to detect naevus count related DNA methylation modifications, rather than dividing the dataset into two smaller subsets that would have lower power to detect associations in order to attempt to provide internal replication. Recently, new skin biopsies have been collected in a subset of the same individuals analysed here as part of a follow up study. Future studies could aim to profile the DNA methylomes of these skin biopsies and perform longitudinal analyses for these individuals or provide replication in independent samples.

## 6.5 Future Perspective

Our knowledge of the epigenome and genomic regulation has been exponentially increasing. Advances in new techniques to interrogate the genome and the epigenome are the driving factor behind this. At the same time, the costs of epigenome-wide approaches, array or NGS based, have been decreasing. As more knowledge is gained about the DNA methylome, for example identification of regions in the genome that are potentially more dynamic [41], more approaches can be tai-

lored to interrogate the DNA methylome cost-effectively. The Illumina beadchips for example, have considerably changed the distribution in the genome of CpG sites interrogated from the widely used epigenome-wide beadchip, the 27k beadchip, to their most recent version in 2016, the EPIC beadchip [75]. For genome-wide NGS approaches, the lower costs for second generation sequencing is also enabling an increase in the number of samples analysed by methods such as WGBS, MeDIP-seq, and RRBS. The epigenomic field is rapidly refining its research questions based on the dynamics observed.

In recent years, single-cell omic profiling techniques have become technically possible. Methods have now been developed for detecting single-cell cytosine modifications, histone modifications, DNA accessibility, and chromosome conformation (see Figure 6.1) [306]. The field of single-cell epigenomics is still in its early stage of development, however, this is expected to rapidly progress due to the development and refinement of these new techniques. One of the advantages of single-cell epigenomics is that it will allow more precise correlation between regulatory epigenetic marks and expression. Therefore it could refine the function of different epigenetic marks such as DNA methylation for example.

Structural views of chromosome folding provided by chromosome conformation capture (3C)-based methods reveal the interactions between chromosomal loci. Hi-C coupled with NGS provides a genome-wide spatial interaction map of nuclear organisation and chromosome architecture [307, 308]. This knowledge will be essential to incorporate into functional genomics and epigenomics.

Third generation sequencing has been developed recently that is distinguished by its ability to directly sequence single molecule strands of DNA over tens of kb. Two methods are currently employed, the use of single-molecule real-time sequencing [309] and the use of electrostatic charge while passing through a protein nano-pore [310]. Its application may not only change genomic analysis but most likely also affect how modifications of DNA can be interrogated. If these new

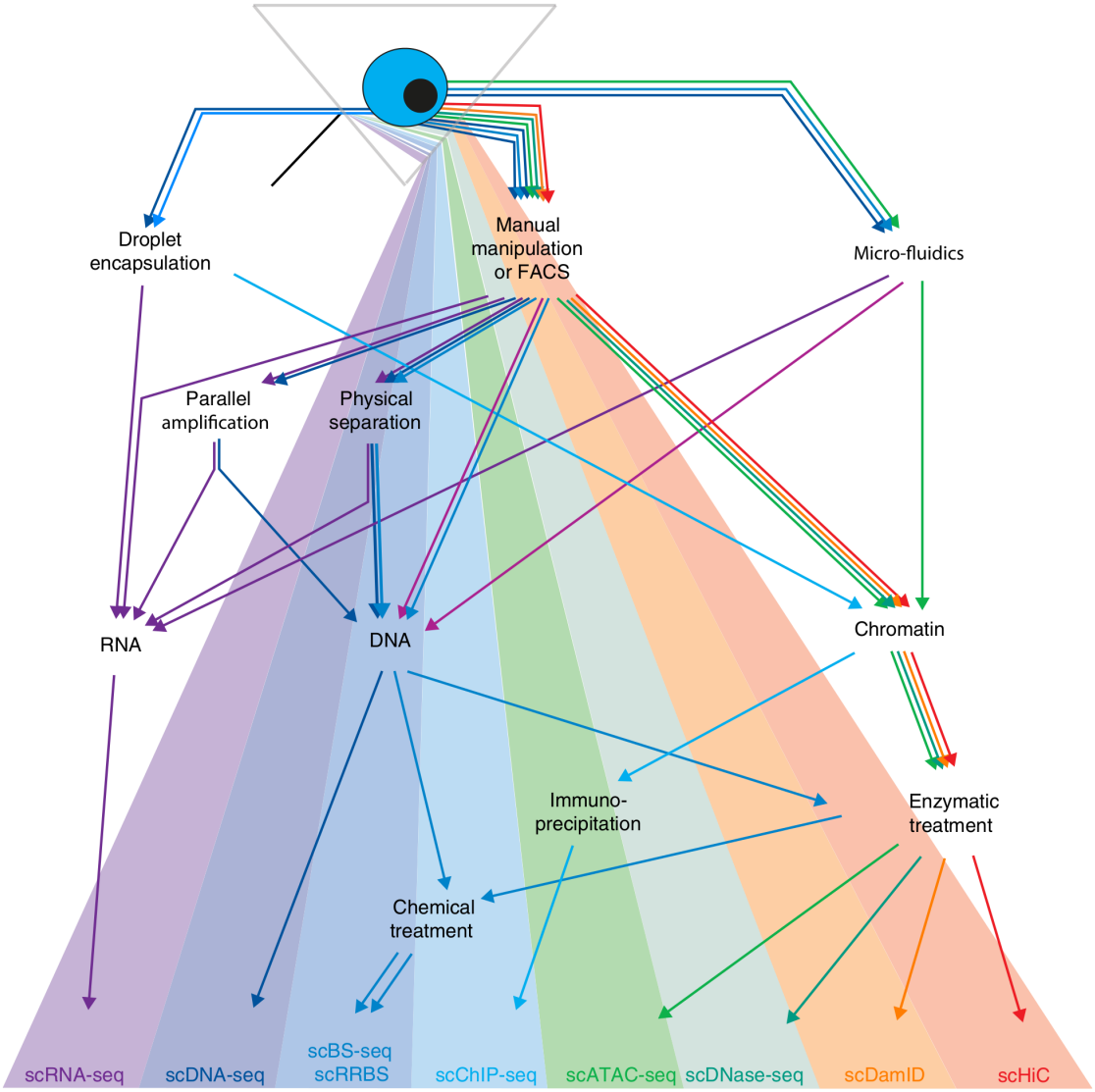



**Figure 6.1: The epigenomics spectrum of single-cell sequencing technologies.** Reproduced from Clark *et al.* [306].

methods can distinguish 5mC or the other rarer cytosine modifications with high sensitivity and specificity, this approach will revolutionise the DNA methylation field by removing the need for an intermediate step to identify DNA modifications. These steps introduce bias, including for example CpG density that influences antibody binding efficiency for MeDIP-seq, bisulphite conversion efficiency for WGBS and RRBS, and sequence dependent restriction enzymes for RRBS. Moreover, longer reads are easier to align to a reference genome, particularly for repetitive regions that are predominantly methylated, and enable epigenetic haplotype variation to be

more accurately assessed. These advances will also strongly benefit single-cell omic approaches.

The EWAS "era" has moved rapidly from assessing single markers, such as DNA methylation and single histone modifications, to integrative studies including multiple facets of epigenomics and/or genomics [311–313]. Ultimately, epigenomics contributes to our understanding of genomic regulation in health and disease. With the application of these new techniques and the knowledge that will arise from the findings, the coming years will bring exciting new understanding of the function of the epigenome. This will potentially enable the future application of epigenomic biomarkers in clinical assessment and highly accurate epigenome modifiers in treatment.

# References


1. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. eng. *Nature Reviews. Genetics* **13,** 484–492. ISSN: 1471-0064 (July 2012).

2. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. eng. *Nature* **518,** 317–330. ISSN: 1476-4687 (Feb. 2015).

3. Bird, A. DNA methylation patterns and epigenetic memory. eng. *Genes & Development* **16,** 6–21. ISSN: 0890-9369 (Jan. 2002).

4. Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. eng. *Nature Reviews. Genetics* **11,** 285–296. ISSN: 1471-0064 (Apr. 2010).

5. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. eng. *Nature Reviews. Genetics* **10,** 295–304. ISSN: 1471-0064 (May 2009).

6. Kowalik, K. M. *et al.* The Paf1 complex represses small-RNA-mediated epigenetic gene silencing. eng. *Nature* **520,** 248–252. ISSN: 1476-4687 (Apr. 2015).

7. Aguilar, C. A. & Craighead, H. G. Micro- and nanoscale devices for the investigation of epigenetics and chromatin dynamics. eng. *Nature Nanotechnology* **8,** 709–718. ISSN: 1748-3395 (Oct. 2013).

8. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. eng. *Nature* **523,** 212–216. ISSN: 1476-4687 (July 2015).

9. Guo, J. U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. en. *Nature Neuroscience* **17,** 215–222. ISSN: 1097-6256 (Feb. 2014).

10. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. eng. *Nature* **462,** 315–322. ISSN: 1476-4687 (Nov. 2009).

11. Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. eng. *Nature Reviews. Genetics* **13,** 7–13. ISSN: 1471-0064 (Jan. 2012).

12. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. eng. *Science (New York, N.Y.)* **324,** 930–935. ISSN: 1095-9203 (May 2009).

13. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. eng. *Nature* **466,** 1129–1133. ISSN: 1476-4687 (Aug. 2010).

14. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. eng. *Science (New York, N.Y.)* **333,** 1300–1303. ISSN: 1095-9203 (Sept. 2011).

15. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. eng. *Science (New York, N.Y.)* **333,** 1303–1307. ISSN: 1095-9203 (Sept. 2011).

16. Wen, L. *et al.* Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. eng. *Genome Biology* **15,** R49. ISSN: 1474-760X (2014).

17. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C. & Greenberg, M. E. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. eng. *Proceedings of the Na-*

*tional Academy of Sciences of the United States of America* **112,** 6800–6806. ISSN: 1091-6490 (June 2015).

18. Szerlong, H. J. & Hansen, J. C. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **89,** 24–34. ISSN: 0829-8211 (Feb. 2011).

19. Bednar, J. *et al.* Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. eng. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 14173–14178. ISSN: 0027-8424 (Nov. 1998).

20. Izzo, A. *et al.* The Genomic Landscape of the Somatic Linker Histone Subtypes H1.1 to H1.5 in Human Cells. *Cell Reports* **3,** 2142–2154. ISSN: 2211-1247 (June 2013).

21. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. eng. *Nature* **473,** 43–49. ISSN: 1476-4687 (May 2011).

22. Rothbart, S. B. & Strahl, B. D. Interpreting thelanguage of histone and DNA modifications. *Biochimica et biophysica acta* **1839,** 627–643. ISSN: 0006-3002 (Aug. 2014).

23. Keung, A. J., Joung, J. K., Khalil, A. S. & Collins, J. J. Chromatin regulation at the frontier of synthetic biology. eng. *Nature Reviews. Genetics* **16,** 159–171. ISSN: 1471-0064 (Mar. 2015).

24. Tessarz, P. & Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. eng. *Nature Reviews. Molecular Cell Biology* **15,** 703–708. ISSN: 1471-0080 (Nov. 2014).

25. Kouzarides, T. Chromatin Modifications and Their Function. English. *Cell* **128,** 693–705. ISSN: 0092-8674, 1097-4172 (Feb. 2007).

26. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. eng. *Nature Methods* **9,** 473–476. ISSN: 1548-7105 (May 2012).

27. Morgan, H. D., Santos, F., Green, K., Dean, W. & Reik, W. Epigenetic reprogramming in mammals. en. *Human Molecular Genetics* **14,** R47–R58. ISSN: 0964-6906, 1460-2083 (Apr. 2005).

28. Lee, H., Hore, T. & Reik, W. Reprogramming the Methylome: Erasing Memory and Creating Diversity. *Cell Stem Cell* **14,** 710–719. ISSN: 1934-5909 (June 2014).

29. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. eng. *Nature* **447,** 425–432. ISSN: 1476-4687 (May 2007).

30. Kota, S. K. & Feil, R. Epigenetic transitions in germ cell development and meiosis. eng. *Developmental Cell* **19,** 675–686. ISSN: 1878-1551 (Nov. 2010).

31. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. en. *Nature Genetics* **48,** 253–264. ISSN: 1061-4036 (Mar. 2016).

32. Horvath, S. DNA methylation age of human tissues and cell types. en. *Genome Biology* **14,** R115. ISSN: 1465-6906 (Oct. 2013).

33. Talens, R. P. *et al.* Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. eng. *Aging Cell* **11,** 694–703. ISSN: 1474-9726 (Aug. 2012).

34. Wong, C. C. Y. *et al.* A longitudinal study of epigenetic variation in twins. eng. *Epigenetics* **5,** 516–526. ISSN: 1559-2308 (Aug. 2010).

35. Rakyan, V. K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. eng. *Genome Research* **20,** 434–439. ISSN: 1549-5469 (Apr. 2010).

36. Teschendorff, A. E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. eng. *Genome Research* **20,** 440–446. ISSN: 1549-5469 (Apr. 2010).

37. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. en. *Nature Reviews Genetics* **13,** 97–109. ISSN: 1471-0056 (Feb. 2012).

38. Weaver, I. C. G. *et al.* Epigenetic programming by maternal behavior. eng. *Nature Neuroscience* **7,** 847–854. ISSN: 1097-6256 (Aug. 2004).

39. Anderson, O. S. *et al.* Epigenetic responses following maternal dietary exposure to physiologically relevant levels of bisphenol A. eng. *Environmental and Molecular Mutagenesis* **53,** 334–342. ISSN: 1098-2280 (June 2012).

40. Tobi, E. W. *et al.* DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. eng. *Human Molecular Genetics* **18,** 4046–4053. ISSN: 1460-2083 (Nov. 2009).

41. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. en. *Nature* **500,** 477–481. ISSN: 0028-0836 (Aug. 2013).

42. Sved, J. & Bird, A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. eng. *Proceedings of the National Academy of Sciences of the United States of America* **87,** 4692–4696. ISSN: 0027-8424 (June 1990).

43. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. eng. *Genes & Development* **25,** 1010–1022. ISSN: 1549-5477 (May 2011).

44. Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. eng. *Cell* **145,** 773–786. ISSN: 1097-4172 (May 2011).

45. Illingworth, R. S. *et al.* Orphan CpG islands identify numerous conserved promoters in the mammalian genome. eng. *PLoS genetics* **6,** e1001134. ISSN: 1553-7404 (Sept. 2010).

46. Illingworth, R. *et al.* A novel CpG island set identifies tissue-specific methylation at developmental gene loci. eng. *PLoS biology* **6,** e22. ISSN: 1545-7885 (Jan. 2008).

47. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. eng. *Nature* **466,** 253–257. ISSN: 1476-4687 (July 2010).

48. Jones, P. A. & Liang, G. Rethinking how DNA methylation patterns are maintained. eng. *Nature Reviews. Genetics* **10,** 805–811. ISSN: 1471-0064 (Nov. 2009).

49. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. eng. *Nature Genetics* **39,** 457–466. ISSN: 1061-4036 (Apr. 2007).

50. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. eng. *Nature Genetics* **42,** 1093–1100. ISSN: 1546-1718 (Dec. 2010).

51. Schübeler, D. Function and information content of DNA methylation. en. *Nature* **517,** 321–326. ISSN: 0028-0836 (Jan. 2015).

52. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. en. *Nature Genetics* **41,** 178–186. ISSN: 1061-4036 (Feb. 2009).

53. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. en. *Nature Genetics* **43,** 768–775. ISSN: 1061-4036 (Aug. 2011).

54. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. en. *Nature Genetics* **41,** 1350–1353. ISSN: 1061-4036 (Dec. 2009).

55. Hodges, E. *et al.* Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. eng. *Molecular Cell* **44,** 17–28. ISSN: 1097-4164 (Oct. 2011).

56. Bock, C. Analysing and interpreting DNA methylation data. en. *Nature Reviews Genetics* **13,** 705–719. ISSN: 1471-0056 (Oct. 2012).

57. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. eng. *Nature* **480,** 490–495. ISSN: 1476-4687 (Dec. 2011).

58. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. en. *Nature* **461,** 747–753. ISSN: 0028-0836 (Oct. 2009).

59. Feinberg AP & Fallin M. EPigenetics at the crossroads of genes and the environment. *JAMA* **314,** 1129–1130. ISSN: 0098-7484 (Sept. 2015).

60. Richards, E. J. Inherited epigenetic variation–revisiting soft inheritance. eng. *Nature Reviews. Genetics* **7,** 395–401. ISSN: 1471-0056 (May 2006).

61. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. eng. *Nature* **490,** 267–272. ISSN: 1476-4687 (Oct. 2012).

62. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. en. *Nature Reviews Genetics* **11,** 446–450. ISSN: 1471-0056 (June 2010).

63. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. eng. *Nature Biotechnology* **31,** 142–147. ISSN: 1546-1696 (Feb. 2013).

64. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. en. *Nature Reviews Cancer* **11,** 726–734. ISSN: 1474-175X (Oct. 2011).

65. Huynh, J. L. *et al.* Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. eng. *Nature Neuroscience* **17,** 121–130. ISSN: 1546-1726 (Jan. 2014).

66. Dick, K. J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *The Lancet* **383,** 1990–1998. ISSN: 0140-6736 (June 2014).

67. Bell, J. T. *et al.* Differential methylation of the TRPA1 promoter in pain sensitivity. eng. *Nature Communications* **5,** 2978. ISSN: 2041-1723 (2014).

68. Elliott, H. R. *et al.* Differences in smoking associated DNA methylation patterns in South Asians and Europeans. en. *Clinical Epigenetics* **6,** 4. ISSN: 1868-7083 (Feb. 2014).

69. Zeilinger, S. *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. eng. *PloS One* **8,** e63812. ISSN: 1932-6203 (2013).

70. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. eng. *Environmental Health Perspectives* **120,** 1425–1431. ISSN: 1552-9924 (Oct. 2012).

71. Schones, D. E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. en. *Nature Reviews Genetics* **9,** 179–191. ISSN: 1471-0056 (Mar. 2008).

72. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. eng. *Nature Genetics* **37,** 853–862. ISSN: 1061-4036 (Aug. 2005).

73. Nair, S. S. *et al.* Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* **6,** 34–44. ISSN: 1559-2294 (Jan. 2011).

74. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. en. *Nature Reviews Genetics* **11,** 191–203. ISSN: 1471-0056 (Feb. 2010).

75. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8,** 389–399. ISSN: 1750-1911 (Dec. 2015).

76. Nestor, C. E. *et al.* Tissue-type is a major modifier of the 5-hydroxymethylcytosine content of human genes. en. *Genome Research,* gr.126417.111. ISSN: 1088-9051, 1549-5469 (Nov. 2011).

77. Booth, M. J. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. en. *Science* **336,** 934–937. ISSN: 0036-8075, 1095-9203 (May 2012).

78. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. eng. *Nature Reviews. Genetics* **16,** 716–726. ISSN: 1471-0064 (Dec. 2015).

79. Bauer, M. *et al.* A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. en. *Clinical Epigenetics* **7,** 1–11. ISSN: 1868-7075, 1868-7083 (Aug. 2015).

80. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. en. *BMC Bioinformatics* **13,** 86. ISSN: 1471-2105 (May 2012).

81. Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in

pathways to disease. *International Journal of Epidemiology* **41,** 161–176. ISSN: 0300-5771 (Feb. 2012).

82. Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling molecular relationships with a causal inference test. *BMC Genetics* **10,** 23. ISSN: 1471-2156 (2009).

83. Van Dongen, J., Slagboom, P. E., Draisma, H. H. M., Martin, N. G. & Boomsma, D. I. The continuing value of twin studies in the omics era. eng. *Nature Reviews. Genetics* **13,** 640–653. ISSN: 1471-0064 (Sept. 2012).

84. Hall, J. G. Twinning. eng. *Lancet (London, England)* **362,** 735–743. ISSN: 1474-547X (Aug. 2003).

85. Hall, J. G. Twins and twinning. eng. *American Journal of Medical Genetics* **61,** 202–204. ISSN: 0148-7299 (Jan. 1996).

86. Kendler, K. S. & Prescott, C. A. A population-based twin study of lifetime major depression in men and women. eng. *Archives of General Psychiatry* **56,** 39–44. ISSN: 0003-990X (Jan. 1999).

87. MacGregor, A. J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. eng. *Arthritis and Rheumatism* **43,** 30–37. ISSN: 0004-3591 (Jan. 2000).

88. Sadovnick, A. D. *et al.* A population-based study of multiple sclerosis in twins: update. eng. *Annals of Neurology* **33,** 281–285. ISSN: 0364-5134 (Mar. 1993).

89. Bogdanos, D. P. *et al.* Twin studies in autoimmune disease: genetics, gender and environment. eng. *Journal of Autoimmunity* **38,** J156–169. ISSN: 1095-9157 (May 2012).

90. Lichtenstein, P. *et al.* Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England Journal of Medicine* **343,** 78–85. ISSN: 0028-4793 (July 2000).

91. Manolio, T. A. Bringing genome-wide association findings into clinical use. eng. *Nature Reviews. Genetics* **14,** 549–558. ISSN: 1471-0064 (Aug. 2013).

92. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. eng. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 10604–10609. ISSN: 0027-8424 (July 2005).

93. Kaminsky, Z. A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. en. *Nature Genetics* **41,** 240–245. ISSN: 1061-4036 (Feb. 2009).

94. Bell, J. T. & Saffery, R. The value of twins in epigenetic epidemiology. eng. *International Journal of Epidemiology* **41,** 140–150. ISSN: 1464-3685 (Feb. 2012).

95. Rakyan, V. K. *et al.* Identification of Type 1 Diabetes–Associated DNA Methylation Variable Positions That Precede Disease Diagnosis. *PLOS Genet* **7,** e1002300. ISSN: 1553-7404 (Sept. 2011).

96. Yuan, W. *et al.* An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. en. *Nature Communications* **5,** 5719 (Dec. 2014).

97. Kuratomi, G. *et al.* Aberrant DNA methylation associated with bipolar disorder identified from discordant monozygotic twins. eng. *Molecular Psychiatry* **13,** 429–441. ISSN: 1476-5578 (Apr. 2008).

98. Javierre, B. M. *et al.* Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. eng. *Genome Research* **20,** 170–179. ISSN: 1549-5469 (Feb. 2010).

99. Selmi, C. *et al.* X chromosome gene methylation in peripheral lymphocytes from monozygotic twins discordant for scleroderma. eng. *Clinical and Experimental Immunology* **169,** 253–262. ISSN: 1365-2249 (Sept. 2012).

100. Nguyen, A., Rauch, T. A., Pfeifer, G. P. & Hu, V. W. Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. eng. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **24,** 3036–3051. ISSN: 1530-6860 (Aug. 2010).

101. Wong, C. C. Y. *et al.* Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. en. *Molecular Psychiatry* **19,** 495–503. ISSN: 1359-4184 (Apr. 2014).

102. Dempster, E. L. *et al.* Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder. eng. *Human Molecular Genetics* **20,** 4786–4796. ISSN: 1460-2083 (Dec. 2011).

103. Allione, A. *et al.* Novel Epigenetic Changes Unveiled by Monozygotic Twins Discordant for Smoking Habits. *PLOS ONE* **10,** e0128265. ISSN: 1932-6203 (June 2015).

104. Pietiläinen, K. H. *et al.* DNA methylation and gene expression patterns in adipose tissue differ significantly within young adult monozygotic BMI-discordant twin pairs. eng. *International Journal of Obesity (2005)* **40,** 654–661. ISSN: 1476-5497 (Apr. 2016).

105. Ollikainen, M. *et al.* Genome-wide blood DNA methylation alterations at regulatory elements and heterochromatic regions in monozygotic twins discordant for obesity and liver fat. en. *Clinical Epigenetics* **7,** 1–13. ISSN: 1868-7075, 1868-7083 (Apr. 2015).

106. Tsai, P.-C. *et al.* DNA Methylation Changes in the IGF1R Gene in Birth Weight Discordant Adult Monozygotic Twins. *Twin Research and Human Genetics* **18,** 635–646. ISSN: 1839-2628 (Dec. 2015).

107. Willemsen, G. *et al.* The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Research and Human Genetics* **18,** 762–771. ISSN: 1839-2628 (Dec. 2015).

108. all International Agency for Research on Cancer. *World Cancer Report 2014* ISBN: 978-92-832-0429-9. <`www.iarc.fr/en/publications/books/wcr/wcr-order.php`> (WHO Press, Feb. 2014).

109. Gulland, A. Global cancer prevalence is growing at "alarming pace," says WHO. en. *BMJ* **348,** g1338–g1338. ISSN: 1756-1833 (Feb. 2014).

110. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70. ISSN: 0092-8674 (Jan. 2000).

111. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. eng. *Cell* **144,** 646–674. ISSN: 1097-4172 (Mar. 2011).

112. You, J. & Jones, P. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell* **22,** 9–20. ISSN: 1535-6108 (July 2012).

113. Klein, A. P. *et al.* Prospective Risk of Pancreatic Cancer in Familial Pancreatic Cancer Kindreds. en. *Cancer Research* **64,** 2634–2638. ISSN: 0008-5472, 1538-7445 (Apr. 2004).

114. Casadei, S. *et al.* Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. eng. *Cancer Research* **71,** 2222–2229. ISSN: 1538-7445 (Mar. 2011).

115. Geary, J. *et al.* Gene-related cancer spectrum in families with hereditary non-polyposis colorectal cancer (HNPCC). eng. *Familial Cancer* **7,** 163–172. ISSN: 1389-9600 (2008).

116. Garber, J. E. & Offit, K. Hereditary cancer predisposition syndromes. eng. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **23,** 276–292. ISSN: 0732-183X (Jan. 2005).

117. GLOBOCAN 2012 (IARC). *GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012* <`http://globocan.iarc.fr/Pages/Map.aspx`> (visited on 06/29/2016).

118. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. eng. *Nature Reviews. Genetics* **8,** 286–298. ISSN: 1471-0056 (Apr. 2007).

119. Bert, S. A. *et al.* Regional activation of the cancer genome by long-range epigenetic remodeling. eng. *Cancer Cell* **23,** 9–22. ISSN: 1878-3686 (Jan. 2013).

120. Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. eng. *Nature Reviews. Genetics* **7,** 21–33. ISSN: 1471-0056 (Jan. 2006).

121. Dawson, M. A., Kouzarides, T. & Huntly, B. J. P. Targeting epigenetic readers in cancer. eng. *The New England Journal of Medicine* **367,** 647–657. ISSN: 1533-4406 (Aug. 2012).

122. Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. eng. *Genome Research* **22,** 246–258. ISSN: 1549-5469 (Feb. 2012).

123. Berman, B. P. *et al.* Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. eng. *Nature Genetics* **44,** 40–46. ISSN: 1546-1718 (Jan. 2012).

124. Timp, W. & Feinberg, A. P. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. eng. *Nature Reviews. Cancer* **13,** 497–510. ISSN: 1474-1768 (July 2013).

125. García, P. *et al.* Promoter methylation profile in preneoplastic and neoplastic gallbladder lesions. eng. *Molecular Carcinogenesis* **48,** 79–89. ISSN: 1098-2744 (Jan. 2009).

126. Baylin, S. B. & Ohm, J. E. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? eng. *Nature Reviews. Cancer* **6,** 107–116. ISSN: 1474-175X (Feb. 2006).

127. Noreen, F. *et al.* Modulation of Age- and Cancer-Associated DNA Methylation Change in the Healthy Colon by Aspirin and Lifestyle. en. *Journal of the National Cancer Institute* **106,** dju161. ISSN: 0027-8874, 1460-2105 (July 2014).

128. Heyn, H. & Esteller, M. DNA methylation profiling in the clinic: applications and challenges. en. *Nature Reviews Genetics* **13,** 679–692. ISSN: 1471-0056 (Oct. 2012).

129. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. en. *Proceedings of the National Academy of Sciences* **112,** E5503–E5512. ISSN: 0027-8424, 1091-6490 (Oct. 2015).

130. Church, T. R. *et al.* Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. en. *Gut* **63,** 317–325. ISSN: , 1468-3288 (Feb. 2014).

131. Yi, J. M. *et al.* Novel Methylation Biomarker Panel for the Early Detection of Pancreatic Cancer. en. *Clinical Cancer Research* **19,** 6544–6555. ISSN: 1078-0432, 1557-3265 (Dec. 2013).

132. Jatkoe, T. A. *et al.* A urine-based methylation signature for risk stratification within low-risk prostate cancer. en. *British Journal of Cancer* **112,** 802–808. ISSN: 0007-0920 (Mar. 2015).

133. Hubers, A. J. *et al.* DNA hypermethylation analysis in sputum for the diagnosis of lung cancer: training validation set approach. en. *British Journal of Cancer* **112,** 1105–1113. ISSN: 0007-0920 (Mar. 2015).

134. Brennan, K. *et al.* Intragenic ATM Methylation in Peripheral Blood DNA as a Biomarker of Breast Cancer Risk. en. *Cancer Research* **72,** 2304–2313. ISSN: 0008-5472, 1538-7445 (May 2012).

135. Steegenga, W. T. *et al.* Genome-wide age-related changes in DNA methylation and gene expression in human PBMCs. eng. *Age (Dordrecht, Netherlands)* **36,** 9648. ISSN: 1574-4647 (June 2014).

136. Anjum, S. *et al.* A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival. *Genome Medicine* **6,** 47. ISSN: 1756-994X (June 2014).

137. Heyn, H. *et al.* DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. en. *Carcinogenesis* **34,** 102–108. ISSN: 0143-3334, 1460-2180 (Jan. 2013).

138. Lim, U. *et al.* Genomic Methylation of Leukocyte DNA in Relation to Colorectal Adenoma Among Asymptomatic Women. *Gastroenterology* **134,** 47–55. ISSN: 0016-5085 (Jan. 2008).

139. Langevin, S. M. *et al.* Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. *Epigenetics* **9,** 884–895. ISSN: 1559-2294 (June 2014).

140. Teschendorff, A. E. *et al.* An Epigenetic Signature in Peripheral Blood Predicts Active Ovarian Cancer. *PLoS ONE* **4,** e8274 (Dec. 2009).

141. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Research and Human Genetics* **16,** 144–149. ISSN: 1839-2628 (Feb. 2013).

142. Andrew, T. *et al.* Are Twins and Singletons Comparable? A Study of Disease-related and Lifestyle Characteristics in Adult Women. *Twin Research and Human Genetics* **4,** 464–477. ISSN: 1839-2628 (Dec. 2001).

143. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. eng. *American Journal of Human Genetics* **93,** 876–890. ISSN: 1537-6605 (Nov. 2013).

144. WHO. *WHO | International Classification of Diseases (ICD)* <`http://www.who.int/classifications/icd/en/`> (visited on 05/19/2016).

145. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. eng. *PLoS genetics* **7,** e1002003. ISSN: 1553-7404 (2011).

146. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. eng. *Nature Genetics* **44,** 1084–1089. ISSN: 1546-1718 (Oct. 2012).

147. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3,** 771–784. ISSN: 1750-1911 (Dec. 2011).

148. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. eng. *Epigenetics* **6,** 692–702. ISSN: 1559-2308 (June 2011).

149. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. en. *Bioinformatics* **30,** 1363–1369. ISSN: 1367-4803, 1460-2059 (May 2014).

150. Butcher, L. *Illumina450ProbeVariants.db: Annotation Package combining variant data from 1000 Genomes Project for Illumina HumanMethylation450 Bead Chip probes. R package version 1.3.1* 2013.

151. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. en. *Nature* **491,** 56–65. ISSN: 0028-0836 (Nov. 2012).

152. Schalkwyk, L. C. *et al. wateRmelon: Illumina 450 methylation array normalization and metrics* 2013.

153. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* **81,** 559–575. ISSN: 0002-9297 (Sept. 2007).

154. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. eng. *Bioinformatics (Oxford, England)* **29,** 189–196. ISSN: 1367-4811 (Jan. 2013).

155. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. en. *Genome Biology* **15,** 503. ISSN: 1465-6906 (Dec. 2014).

156. Butcher, L. M. & Beck, S. AutoMeDIP-seq: a high-throughput, whole genome, DNA methylation assay. eng. *Methods (San Diego, Calif.)* **52,** 223–231. ISSN: 1095-9130 (Nov. 2010).

157. Chavez, L. *et al.* Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. en. *Genome Research* **20,** 1441–1450. ISSN: 1088-9051, 1549-5469 (Oct. 2010).

158. Davies, M. N. *et al.* Hypermethylation in the ZBTB20 gene is associated with major depressive disorder. *Genome Biology* **15,** R56. ISSN: 1474-760X (2014).

159. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. eng. *PloS One* **7,** e30619. ISSN: 1932-6203 (2012).

160. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. eng. *Bioinformatics (Oxford, England)* **25,** 1754–1760. ISSN: 1367-4811 (July 2009).

161. Small, K. S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. eng. *Nature Genetics* **43,** 561–564. ISSN: 1546-1718 (June 2011).

162. Hanash, S. M., Baik, C. S. & Kallioniemi, O. Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer. en. *Nature Reviews Clinical Oncology* **8,** 142–150. ISSN: 1759-4774 (Mar. 2011).

163. Ogino, S. *et al.* Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease. en. *Modern Pathology* **26,** 465–484. ISSN: 0893-3952 (Apr. 2013).

164. Breitling, L. P., Yang, R., Korn, B., Burwinkel, B. & Brenner, H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. eng. *American Journal of Human Genetics* **88,** 450–457. ISSN: 1537-6605 (Apr. 2011).

165. Wan, E. S. *et al.* Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. eng. *Human Molecular Genetics* **21,** 3073–3082. ISSN: 1460-2083 (July 2012).

166. Panni, T. *et al.* A Genome-Wide Analysis of DNA Methylation and Fine Particulate Matter Air Pollution in Three Study Populations: KORA F3, KORA F4, and the Normative Aging Study. ENG. *Environmental Health Perspectives.* ISSN: 1552-9924. doi:10.1289/ehp.1509966 (Jan. 2016).

167. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. en. *Nature Genetics* **45,** 1113–1120. ISSN: 1061-4036 (Oct. 2013).

168. Witte, T., Plass, C. & Gerhauser, C. Pan-cancer patterns of DNA methylation. eng. *Genome Medicine* **6,** 66. ISSN: 1756-994X (2014).

169. Jandaghi, P., Hoheisel, J. D. & Riazalhosseini, Y. GHSRhypermethylation: a promising pan-cancer marker. *Cell Cycle* **0,** 00–00. ISSN: 1538-4101 (2015).

170. Akbani, R. *et al.* A pan-cancer proteomic perspective on The Cancer Genome Atlas. eng. *Nature Communications* **5,** 3887. ISSN: 2041-1723 (2014).

171. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. en. *Nature* **502,** 333–339. ISSN: 0028-0836 (Oct. 2013).

172. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. en. *Nature Genetics* **45,** 1134–1140. ISSN: 1061-4036 (Oct. 2013).

173. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. eng. *Bioinformatics (Oxford, England)* **30,** 1431–1439. ISSN: 1367-4811 (May 2014).

174. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. eng. *Nature Methods* **11,** 309–311. ISSN: 1548-7105 (Mar. 2014).

175. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. en. *Nature Methods* **13,** 443–445. ISSN: 1548-7091 (May 2016).

176. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. en. *International Journal of Epidemiology* **41,** 200–209. ISSN: 0300-5771, 1464-3685 (Feb. 2012).

177. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. eng. *Journal of Molecular Biology* **196,** 261–282. ISSN: 0022-2836 (July 1987).

178. Ernst, J. *et al.* Systematic analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49. ISSN: 0028-0836 (May 2011).

179. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. eng. *Nature Biotechnology* **28,** 817–825. ISSN: 1546-1696 (Aug. 2010).

180. Willemsen, G. *et al.* The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. *Twin Research and Human Genetics* **13,** 231–245. ISSN: 1839-2628 (June 2010).

181. Gordon, L. *et al.* Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. en. *Genome Research* **22,** 1395–1406. ISSN: 1088-9051, 1549-5469 (Aug. 2012).

182. Bell, J. T. *et al.* Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population. *PLoS Genet* **8,** e1002629 (Apr. 2012).

183. Meng, Q. *et al.* SASH1 regulates proliferation, apoptosis, and invasion of osteosarcoma cell. en. *Molecular and Cellular Biochemistry* **373,** 201–210. ISSN: 0300-8177, 1573-4919 (Oct. 2012).

184. Zeller, C. *et al.* SASH1: a candidate tumor suppressor gene on chromosome 6q24.3 is downregulated in breast cancer. en. *Oncogene* **22,** 2972–2983. ISSN: 0950-9232 (2003).

185. Rimkus, C. *et al.* Prognostic significance of downregulated expression of the candidate tumour suppressor gene SASH1 in colon cancer. en. *British Journal of Cancer* **95,** 1419–1423. ISSN: 0007-0920 (Oct. 2006).

186. Leconet, W. *et al.* Preclinical validation of AXL receptor as a target for antibody-based pancreatic cancer immunotherapy. en. *Oncogene* **33,** 5405–5414. ISSN: 0950-9232 (Nov. 2014).

187. Bansal, N., Mishra, P. J., Stein, M., DiPaola, R. S. & Bertino, J. R. Axl receptor tyrosine kinase is up-regulated in metformin resistant prostate cancer cells. en. *Oncotarget* **6,** 15321 (June 2015).

188. Martinelli, E. *et al.* AXL is an oncotarget in human colorectal cancer. eng. *Oncotarget* **6,** 23281–23296. ISSN: 1949-2553 (Sept. 2015).

189. Cheung, N.-K. V. & Dyer, M. A. Neuroblastoma: developmental biology, cancer genomics and immunotherapy. en. *Nature Reviews Cancer* **13,** 397–411. ISSN: 1474-175X (June 2013).

190. Cicek, M. S. *et al.* Epigenome-wide ovarian cancer analysis identifies a methylation profile differentiating clear-cell histology with epigenetic silencing of the HERG K+ channel. en. *Human Molecular Genetics* **22,** 3038–3047. ISSN: 0964-6906, 1460-2083 (Aug. 2013).

191. Bonora, E., Evangelisti, C., Bonichon, F., Tallini, G. & Romeo, G. Novel germline variants identified in the inner mitochondrial membrane transporter TIMM44 and their role in predisposition to oncocytic thyroid carcinomas. en. *British Journal of Cancer* **95,** 1529–1536. ISSN: 0007-0920 (Oct. 2006).

192. Keita, M. *et al.* Global methylation profiling in serous ovarian cancer is indicative for distinct aberrant DNA methylation signatures associated with tumor aggressiveness and disease progression. *Gynecologic Oncology* **128,** 356–363. ISSN: 0090-8258 (Feb. 2013).

193. Specht, K. *et al.* Expression profiling identifies genes that predict recurrence of breast cancer after adjuvant CMF-based chemotherapy. en. *Breast Cancer Research and Treatment* **118,** 45–56. ISSN: 0167-6806, 1573-7217 (Oct. 2008).

194. Marsit, C. J. *et al.* DNA Methylation Array Analysis Identifies Profiles of Blood-Derived DNA Methylation Associated With Bladder Cancer. en. *Journal of Clinical Oncology* **29,** 1133–1139. ISSN: 0732-183X, 1527-7755 (Mar. 2011).

195. Brooks, A. N. *et al.* A Pan-Cancer Analysis of Transcriptome Changes Associated with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events. *PLoS ONE* **9,** e87361 (Jan. 2014).

196. Çalışkan, M., Pritchard, J. K., Ober, C. & Gilad, Y. The Effect of Freeze-Thaw Cycles on Gene Expression Levels in Lymphoblastoid Cell Lines. *PLoS ONE* **9,** e107166 (Sept. 2014).

197. Grafodatskaya, D. *et al.* EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. *Genomics* **95,** 73–83. ISSN: 0888-7543 (Feb. 2010).

198. Joehanes, R. *et al.* Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the Framingham Heart Study. en. *Physiological Genomics* **44,** 59–75. ISSN: 1094-8341, 1531-2267 (Jan. 2012).

199. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. en. *Nature Reviews Genetics* **9,** 465–476. ISSN: 1471-0056 (June 2008).

200. Bessette, D. C., Qiu, D. & Pallen, C. J. PRL PTPs: mediators and markers of cancer progression. en. *Cancer and Metastasis Reviews* **27,** 231–252. ISSN: 0167-7659, 1573-7233 (Jan. 2008).

201. Dumaual, C. M. *et al.* Tissue-specific alterations of PRL-1 and PRL-2 expression in cancer. *American Journal of Translational Research* **4,** 83–101. ISSN: 1943-8141 (Jan. 2012).

202. Hardy, S., Wong, N. N., Muller, W. J., Park, M. & Tremblay, M. L. Overexpression of the protein tyrosine phosphatase PRL-2 correlates with breast tumor formation and progression. eng. *Cancer Research* **70,** 8959–8967. ISSN: 1538-7445 (Nov. 2010).

203. Liu, Y. *et al.* GeMes, Clusters of DNA Methylation under Genetic Control, Can Inform Genetic and Epigenetic Analysis of Disease. *The American Journal of Human Genetics* **94,** 485–495. ISSN: 0002-9297 (Apr. 2014).

204. Tsai, P.-C. & Bell, J. T. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. en. *International Journal of Epidemiology,* dyv041. ISSN: 0300-5771, 1464-3685 (May 2015).

205. Cohen, J. *Power Analysis for the behavioral Sciences.* 2nd ().

206. Cancer Research UK. *Data Table: Cancer cases and rates by country in the UK* <`http://publications.cancerresearchuk.org/cancerstats/statsincidence`> (visited on 06/06/2016).

207. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. eng. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 10869–10874. ISSN: 0027-8424 (Sept. 2001).

208. Gatza, M. L. *et al.* A pathway-based classification of human breast cancer. eng. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 6994–6999. ISSN: 1091-6490 (Apr. 2010).

209. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. en. *Nature* **534,** 47–54. ISSN: 0028-0836 (June 2016).

210. Peto, J. & Mack, T. M. High constant incidence in twins and other relatives of women with breast cancer. eng. *Nature Genetics* **26,** 411–414. ISSN: 1061-4036 (Dec. 2000).

211. Möller, S. *et al.* The Heritability of Breast Cancer among Women in the Nordic Twin Study of Cancer. en. *Cancer Epidemiology Biomarkers & Prevention* **25,** 145–150. ISSN: 1055-9965, 1538-7755 (Jan. 2016).

212. King, M.-C., Marks, J. H. & Mandell, J. B. Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. en. *Science* **302,** 643–646. ISSN: 0036-8075, 1095-9203 (Oct. 2003).

213. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. en. *Nature Genetics* **45,** 353–361. ISSN: 1061-4036 (Apr. 2013).

214. Antoniou, A. C. *et al.* Common Breast Cancer-Predisposition Alleles Are Associated with Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. *The American Journal of Human Genetics* **82,** 937–948. ISSN: 0002-9297 (Apr. 2008).

215. *Breast cancer incidence (invasive) statistics* May 2015. <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive> (visited on 06/07/2016).

216. Bleyer, A. & Welch, H. G. Effect of Three Decades of Screening Mammography on Breast-Cancer Incidence. *New England Journal of Medicine* **367,** 1998–2005. ISSN: 0028-4793 (Nov. 2012).

217. Widschwendter, M. *et al.* Epigenotyping in Peripheral Blood Cell DNA and Breast Cancer Risk: A Proof of Principle Study. *PLoS ONE* **3,** e2656 (July 2008).

218. Flanagan, J. M. *et al.* Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients. en. *Human Molecular Genetics* **18,** 1332–1342. ISSN: 0964-6906, 1460-2083 (Apr. 2009).

219. Xu, Z. *et al.* Epigenome-wide association study of breast cancer using prospectively collected sister study samples. eng. *Journal of the National Cancer Institute* **105,** 694–700. ISSN: 1460-2105 (May 2013).

220. Severi, G. *et al.* Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. eng. *Breast Cancer Research and Treatment* **148,** 665–673. ISSN: 1573-7217 (Dec. 2014).

221. Van Veldhoven, K. *et al.* Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. *Clinical Epigenetics* **7,** 67. ISSN: 1868-7083 (2015).

222. Yang, R. *et al.* DNA methylation array analyses identified breast cancer-associated HYAL2 methylation in peripheral blood. en. *International Journal of Cancer* **136,** 1845–1855. ISSN: 1097-0215 (Apr. 2015).

223. Roos, L. *et al.* Integrative DNA methylome analysis of pan-cancer biomarkers in cancer discordant monozygotic twin-pairs. *Clinical Epigenetics* **8,** 7. ISSN: 1868-7083 (2016).

224. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. eng. *Bioinformatics (Oxford, England)* **23,** 1294–1296. ISSN: 1367-4811 (May 2007).

225. Feng, M. *et al.* RASAL2 activates RAC1 to promote triple-negative breast cancer progression. *The Journal of Clinical Investigation* **124,** 5291–5304. ISSN: 0021-9738 (Dec. 2014).

226. Amelio, I., Knight, R. A., Lisitsa, A., Melino, G. & Antonov, A. V. p53MutaGene: an online tool to estimate the effect of p53 mutational status on gene regulation in cancer. en. *Cell Death & Disease* **7,** e2148 (Mar. 2016).

227. Schumacker, P. Reactive Oxygen Species in Cancer: A Dance with the Devil. *Cancer Cell* **27,** 156–157. ISSN: 1535-6108 (Feb. 2015).

228. Harris, I. *et al.* Glutathione and Thioredoxin Antioxidant Pathways Synergize to Drive Cancer Initiation and Progression. *Cancer Cell* **27,** 211–222. ISSN: 1535-6108 (Feb. 2015).

229. Figueroa, J. D. *et al.* Identification of a novel susceptibility locus at 13q34 and refinement of the 20p12.2 region as a multi-signal locus associated with bladder cancer risk in individuals of European ancestry. en. *Human Molecular Genetics,* ddv492. ISSN: 0964-6906, 1460-2083 (Jan. 2016).

230. Kaczkowski, B. *et al.* Transcriptome Analysis of Recurrently Deregulated Genes across Multiple Cancers Identifies New Pan-Cancer Biomarkers. en. *Cancer Research* **76,** 216–226. ISSN: 0008-5472, 1538-7445 (Jan. 2016).

231. Yang, B. *et al.* Methylation Profiling Defines an Extensive Field Defect in Histologically Normal Prostate Tissues Associated with Prostate Cancer. *Neoplasia* **15,** 399–IN13. ISSN: 1476-5586 (Apr. 2013).

232. Wang, S., Robertson, G. P. & Zhu, J. A novel human homologue of Drosophila polycomblike gene is up-regulated in multiple cancers. eng. *Gene* **343,** 69–78. ISSN: 0378-1119 (Dec. 2004).

233. Scelfo, A., Piunti, A. & Pasini, D. The controversial role of the Polycomb group proteins in transcription and cancer: how much do we not understand Polycomb proteins? en. *FEBS Journal* **282,** 1703–1722. ISSN: 1742-4658 (May 2015).

234. Sjöblom, T. *et al.* The Consensus Coding Sequences of Human Breast and Colorectal Cancers. en. *Science* **314,** 268–274. ISSN: 0036-8075, 1095-9203 (Oct. 2006).

235. Fleischer, T. *et al.* Genome-wide DNA methylation profiles in progression to in situand invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biology* **15,** 435. ISSN: 1474-760X (2014).

236. Cheong, S.-M., Choi, H., Hong, B. S., Gho, Y. S. & Han, J.-K. Dab2 is pivotal for endothelial cell migration by mediating VEGF expression in cancer cells. *Experimental Cell Research* **318,** 550–557. ISSN: 0014-4827 (Mar. 2012).

237. Du, L. *et al.* miR-93-directed downregulation of DAB2 defines a novel oncogenic pathway in lung cancer. en. *Oncogene* **33,** 4307–4315. ISSN: 0950-9232 (Aug. 2014).

238. Jacobsen, A. *et al.* Analysis of microRNA-target interactions across diverse cancer types. en. *Nature Structural & Molecular Biology* **20,** 1325–1332. ISSN: 1545-9993 (Nov. 2013).

239. Kent, W. J. *et al.* The Human Genome Browser at UCSC. en. *Genome Research* **12,** 996–1006. ISSN: 1088-9051, 1549-5469 (June 2002).

240. Devlin, B. & Roeder, K. Genomic control for association studies. eng. *Biometrics* **55,** 997–1004. ISSN: 0006-341X (Dec. 1999).

241. Tapper, W. *et al.* Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. en. *Nature Communications* **6,** 6691 (Apr. 2015).

242. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. en. *Nature* **474,** 609–615. ISSN: 0028-0836 (June 2011).

243. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. en. *Genome Research* **22,** 568–576. ISSN: 1088-9051, 1549-5469 (Mar. 2012).

244. Ramakrishna, M. *et al.* Identification of Candidate Growth Promoting Genes in Ovarian Cancer through Integrated Copy Number and Expression Analysis. *PLOS ONE* **5,** e9983. ISSN: 1932-6203 (Apr. 2010).

245. Perryman, L. & Erler, J. T. Brain Cancer Spreads. en. *Science Translational Medicine* **6,** 247fs28–247fs28. ISSN: 1946-6234, 1946-6242 (July 2014).

246. Sayadi, A. *et al.* Functional features of EVI1 and EVI1Δ324 isoforms of MECOM gene in genome-wide transcription regulation and oncogenicity. en. *Oncogene* **35,** 2311–2321. ISSN: 0950-9232 (May 2016).

247. Patel, J. B. *et al.* Control of EVI-1 oncogene expression in metastatic breast cancer cells through microRNA miR-22. en. *Oncogene* **30,** 1290–1301. ISSN: 0950-9232 (Mar. 2011).

248. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. en. *Nature Structural & Molecular Biology* **20,** 1147–1155. ISSN: 1545-9993 (Oct. 2013).

249. Clark, C. *et al.* A Comparison of the Whole Genome Approach of MeDIP-Seq to the Targeted Approach of the Infinium HumanMethylation450 BeadChip ® for Methylome Profiling. *PLOS ONE* **7,** e50233. ISSN: 1932-6203 (Nov. 2012).

250. Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations. en. *Nature Protocols* **7,** 617–636. ISSN: 1754-2189 (Apr. 2012).

251. Garbe, C. & Leiter, U. Melanoma epidemiology and trends. *Clinics in Dermatology. Melanoma and Pigmented Lesions, Part 1* **27,** 3–9. ISSN: 0738-081X (Jan. 2009).

252. Gandini, S. *et al.* Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi. *European Journal of Cancer* **41,** 28–44. ISSN: 0959-8049 (Jan. 2005).

253. Olsen, C. M. *et al.* Nevus density and melanoma risk in women: A pooled analysis to test the divergent pathway hypothesis. en. *International Journal of Cancer* **124,** 937–944. ISSN: 1097-0215 (Feb. 2009).

254. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. en. *CA: A Cancer Journal for Clinicians* **64,** 9–29. ISSN: 1542-4863 (Jan. 2014).

255. Ferlay, J. *et al.* Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. English. *European Journal of Cancer* **49,** 1374–1403. ISSN: 0959-8049, 1879-0852 (Apr. 2013).

256. Weatherhead, S. C., Haniffa, M. & Lawrence, C. M. Melanomas arising from naevi and de novo melanomas — does origin matter? en. *British Journal of Dermatology* **156,** 72–76. ISSN: 1365-2133 (Jan. 2007).

257. Shitara, D. *et al.* Nevus-Associated Melanomas. en. *American Journal of Clinical Pathology* **142,** 485–491. ISSN: 0002-9173, 1943-7722 (Oct. 2014).

258. Purdue, M. P. *et al.* Etiologic and Other Factors Predicting Nevus-Associated Cutaneous Malignant Melanoma. en. *Cancer Epidemiology Biomarkers & Prevention* **14,** 2015–2022. ISSN: 1055-9965, 1538-7755 (Aug. 2005).

259. Cichorek, M., Wachulska, M., Stasiewicz, A. & Tymińska, A. Skin melanocytes: biology and development. *Advances in Dermatology and Allergology/Postpy Dermatologii I Alergologii* **30,** 30–41. ISSN: 1642-395X (Feb. 2013).

260. Mark, G. J., Mihm, M. C., Liteplo, M. G., Reed, R. J. & Clark, W. H. Congenital melanocytic nevi of the small and garment type. Clinical, histologic, and ultrastructural studies. eng. *Human Pathology* **4,** 395–418. ISSN: 0046-8177 (Sept. 1973).

261. Tannous, Z. S., Mihm Jr., M. C., Sober, A. J. & Duncan, L. M. Congenital melanocytic nevi: Clinical and histopathologic features, risk of melanoma, and clinical management. *Journal of the American Academy of Dermatology* **52,** 197–203. ISSN: 0190-9622 (Feb. 2005).

262. Zayour, M. & Lazova, R. Congenital melanocytic nevi. eng. *Clinics in Laboratory Medicine* **31,** 267–280. ISSN: 1557-9832 (June 2011).

263. Zalaudek I, Schmid K, Marghoob AA & et al. Frequency of dermoscopic nevus subtypes by age and body site: A cross-sectional study. *Archives of Dermatology* **147,** 663–670. ISSN: 0003-987X (June 2011).

264. Autier, P. *et al.* Sex Differences in Numbers of Nevi on Body Sites of Young European Children: Implications for the Etiology of Cutaneous Melanoma. en. *Cancer Epidemiology Biomarkers & Prevention* **13,** 2003–2005. ISSN: 1055-9965, 1538-7755 (Dec. 2004).

265. Whiteman, D. C. *et al.* Anatomic site, sun exposure, and risk of cutaneous melanoma. eng. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **24,** 3172–3177. ISSN: 1527-7755 (July 2006).

266. Bataille, V. Melanoma. Shall we move away from the sun and focus more on embryogenesis, body weight and longevity? *Medical Hypotheses* **81,** 846–850. ISSN: 0306-9877 (Nov. 2013).

267. Bulliard, J.-L., De Weck, D., Fisch, T., Bordoni, A. & Levi, F. Detailed site distribution of melanoma and sunlight exposure: aetiological patterns from a Swiss series. eng. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO* **18,** 789–794. ISSN: 0923-7534 (Apr. 2007).

268. Newton, J. A. *et al.* How common is the atypical mole syndrome phenotype in apparently sporadic melanoma? English. *Journal of the American Academy of Dermatology* **29,** 989–996. ISSN: 0190-9622 (Dec. 1993).

269. Bataille, V. *et al.* Nevus Size and Number Are Associated with Telomere Length and Represent Potential Markers of a Decreased Senescence In vivo. en. *Cancer Epidemiology Biomarkers & Prevention* **16,** 1499–1502. ISSN: 1055-9965, 1538-7755 (July 2007).

270. Whiteman, D. C. *et al.* Melanocytic nevi, solar keratoses, and divergent pathways to cutaneous melanoma. eng. *Journal of the National Cancer Institute* **95,** 806–812. ISSN: 1460-2105 (June 2003).

271. Bataille, V. *et al.* Solar keratoses: a risk factor for melanoma but negative association with melanocytic naevi. eng. *International Journal of Cancer* **78,** 8–12. ISSN: 0020-7136 (Sept. 1998).

272. Falchi, M. *et al.* Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. en. *Nature Genetics* **41,** 915–919. ISSN: 1061-4036 (Aug. 2009).

273. Nan, H. *et al.* Genome-wide association study identifies nidogen 1 (NID1) as a susceptibility locus to cutaneous nevi and melanoma risk. en. *Human Molecular Genetics,* ddr154. ISSN: 0964-6906, 1460-2083 (Apr. 2011).

274. Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. eng. *Nature Genetics* **41,** 920–925. ISSN: 1546-1718 (Aug. 2009).

275. Barrett, J. H. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. eng. *Nature Genetics* **43,** 1108–1113. ISSN: 1546-1718 (Nov. 2011).

276. Bataille, V., Snieder, H., MacGregor, A. J., Sasieni, P. & Spector, T. D. Genetics of risk factors for melanoma: an adult twin study of nevi and freckles. eng. *Journal of the National Cancer Institute* **92,** 457–463. ISSN: 0027-8874 (Mar. 2000).

277. Vandiver, A. R. *et al.* Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. en. *Genome Biology* **16,** 80. ISSN: 1465-6906 (Apr. 2015).

278. Bass, J., Dabney, A & Robinson, D. *qvalue: Q-value estimation for false discovery rate control. R package version 2.4.2* 2015. <`https://github.com/jdstorey/qvalue`> (visited on 06/21/2016).

279. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. eng. *Bioinformatics (Oxford, England)* **26,** 2336–2337. ISSN: 1367-4811 (Sept. 2010).

280. Ushach, I. *et al.* METEORIN-LIKE is a cytokine associated with barrier tissues and alternatively activated macrophages. *Clinical Immunology* **156,** 119–127. ISSN: 1521-6616 (Feb. 2015).

281. Freiberger, S. N. *et al.* Ingenol Mebutate Signals via PKC/MEK/ERK in Keratinocytes and Induces Interleukin Decoy Receptors IL1R2 and IL13RA2. en. *Molecular Cancer Therapeutics* **14,** 2132–2142. ISSN: 1535-7163, 1538-8514 (Sept. 2015).

282. Puca, L., Chastagner, P., Meas-Yedid, V., Israël, A. & Brou, C. α-arrestin 1 (ARRDC1) and β-arrestins cooperate to mediate Notch degradation in mammals. en. *J Cell Sci* **126,** 4457–4468. ISSN: 0021-9533, 1477-9137 (Oct. 2013).

283. Pinnix, C. C. *et al.* Active Notch1 Confers a Transformed Phenotype to Primary Human Melanocytes. en. *Cancer research* **69,** 5312. ISSN: 10.1158/0008-5472.CAN-08-3767 (July 2009).

284. Massi, D. *et al.* Evidence for differential expression of Notch receptors and their ligands in melanocytic nevi and cutaneous malignant melanoma. en. *Modern Pathology* **19,** 246–254. ISSN: 0893-3952 (Feb. 2006).

285. Flaherty, K. T., Hodi, F. S. & Bastian, B. C. Mutation-driven drug development in melanoma. en. *Current opinion in oncology* **22,** 178 (May 2010).

286. Bollag, G. *et al.* Vemurafenib: the first drug approved for BRAF-mutant cancer. en. *Nature Reviews Drug Discovery* **11,** 873–886. ISSN: 1474-1776 (Nov. 2012).

287. Nakajima, H. *et al.* Loss of HITS (FAM107B) expression in cancers of multiple organs: tissue microarray analysis. *International Journal of Oncology* **41,** 1347–1357. ISSN: 1019-6439 (Oct. 2012).

288. Jin, S.-G., Xiong, W., Wu, X., Yang, L. & Pfeifer, G. P. The DNA methylation landscape of human melanoma. *Genomics* **106,** 322–330. ISSN: 0888-7543 (Dec. 2015).

289. Koga, Y. *et al.* Genome-wide screen of promoter methylation identifies novel markers in melanoma. eng. *Genome Research* **19,** 1462–1470. ISSN: 1088-9051 (Aug. 2009).

290. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. en. *Nucleic Acids Research* **42,** D1001–D1006. ISSN: 0305-1048, 1362-4962 (Jan. 2014).

291. Mangino, M. *et al.* Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. *Human Molecular Genetics* **21,** 5385–5394. ISSN: 0964-6906 (Dec. 2012).

292. Iles, M. M. *et al.* The effect on melanoma risk of genes previously associated with telomere length. eng. *Journal of the National Cancer Institute* **106.** ISSN: 1460-2105. doi:10.1093/jnci/dju267 (Oct. 2014).

293. Caini, S. *et al.* Telomere length and the risk of cutaneous melanoma and non-melanoma skin cancer: a review of the literature and meta-analysis. eng. *Journal of Dermatological Science* **80,** 168–174. ISSN: 1873-569X (Dec. 2015).

294. Nsengimana, J. *et al.* Independent replication of a melanoma subtype gene signature and evaluation of its prognostic value and biological correlates in a population cohort. *Oncotarget* **6,** 11683–11693. ISSN: 1949-2553 (Mar. 2015).

295. Emuss, V., Garnett, M., Mason, C., Project, T. C. G. & Marais, R. Mutations of C-RAF Are Rare in Human Cancer because C-RAF Has a Low Basal Kinase Activity Compared with B-RAF. en. *Cancer Research* **65,** 9719–9726. ISSN: 0008-5472, 1538-7445 (Nov. 2005).

296. Easwaran, H., Tsai, H.-C. & Baylin, S. Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. *Molecular Cell* **54,** 716–727. ISSN: 1097-2765 (June 2014).

297. Rodríguez, E. *et al.* An Integrated Epigenetic and Transcriptomic Analysis Reveals Distinct Tissue-Specific Patterns of DNA Methylation Associated with Atopic Dermatitis. *Journal of Investigative Dermatology* **134,** 1873–1883. ISSN: 0022-202X (July 2014).

298. Roberson, E. D. *et al.* A subset of methylated CpG sites differentiate psoriatic from normal skin. *The Journal of investigative dermatology* **132,** 583–592. ISSN: 0022-202X (Mar. 2012).

299. Singmann, P. *et al.* Characterization of whole-genome autosomal differences of DNA methylation between men and women. En. *Epigenetics & Chromatin* **8,** 1. ISSN: 1756-8935 (Oct. 2015).

300. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. en. *Nature* **489,** 57–74. ISSN: 0028-0836 (Sept. 2012).

301. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. en. *Nature Biotechnology* **28,** 1045–1048. ISSN: 1087-0156 (Oct. 2010).

302. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. en. *Nature Biotechnology* **30,** 224–226. ISSN: 1087-0156 (Mar. 2012).

303. Mackenbach, J. P., Looman, C. W. & van der Meer, J. B. Differences in the misreporting of chronic conditions, by level of education: the effect on inequalities in prevalence rates. *American Journal of Public Health* **86,** 706–711. ISSN: 0090-0036 (May 1996).

304. Stevens, M. *et al.* Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. en. *Genome Research* **23,** 1541–1553. ISSN: 1088-9051, 1549-5469 (Sept. 2013).

305. Zhang, B. *et al.* Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. en. *Genome Research* **23,** 1522–1540. ISSN: 1088-9051, 1549-5469 (Sept. 2013).

306. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biology* **17,** 72. ISSN: 1474-760X (2016).

307. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. eng. *Methods (San Diego, Calif.)* **58,** 268–276. ISSN: 1095-9130 (Nov. 2012).

308. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)* **326,** 289–293. ISSN: 0036-8075 (Oct. 2009).

309. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. en. *Science* **323,** 133–138. ISSN: 0036-8075, 1095-9203 (Jan. 2009).

310. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. en. *Nature Methods* **7,** 461–465. ISSN: 1548-7091 (June 2010).

311. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. eng. *Nature Reviews. Genetics* **12,** 529–541. ISSN: 1471-0064 (Aug. 2011).

312. Mill, J. & Heijmans, B. T. From promises to practical strategies in epigenetic epidemiology. en. *Nature Reviews Genetics* **14,** 585–594. ISSN: 1471-0056 (Aug. 2013).

313. Birney, E., Smith, G. D. & Greally, J. M. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. *PLOS Genet* **12,** e1006105. ISSN: 1553-7404 (June 2016).

314. Cancer Research UK. *CancerStats Data Tables* <`http://publications.cancerresearchuk.org/publicationformat/data_tables`> (visited on 06/07/2016).

# Appendix A

# Supplementary Figures

**UK CANCER INCIDENCE (2013) BY COUNTRY - FEMALES**
**January 2016**

| Cancer Type | ICD-10 code/s | Number of cases | | | | | Incidence rate per 100,000 | | | | | Top 20 UK Rank (cases) | % of all malignant neoplasms excl NMSC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | England | Wales | Scotland | Northern Ireland | UK | England | Wales | Scotland | Northern Ireland | UK | | |
| Anus | C21 | 646 | 58 | 79 | 14 | 797 | 2.5 | 3.5 | 2.8 | 1.8 | 2.5 | | 0.5% |
| Bladder | C67 | 2,396 | 162 | 259 | 59 | 2,876 | 8.9 | 9.5 | 9.3 | 7.3 | 8.9 | 14 | 1.7% |
| Bone Sarcoma | C40-C41 | 234 | 12 | 22 | 5 | 273 | 0.9 | 0.8 | 0.8 | 0.6 | 0.8 | | 0.2% |
| Bowel | C18-C20 | 14,926 | 936 | 1,735 | 558 | 18,155 | 56.2 | 56.1 | 62.9 | 67.3 | 57.1 | 3 | 10.5% |
| Colon | C18 | 10,763 | 652 | 1,285 | 407 | 13,107 | 40.4 | 38.7 | 46.5 | 49.1 | 41.1 | | 7.6% |
| Rectum and Rectosigmoid Junction | C19-C20 | 4,163 | 284 | 450 | 151 | 5,048 | 15.8 | 17.3 | 16.4 | 18.2 | 16.0 | | 2.9% |
| Brain, Other CNS and Intracranial Tumours | C70-C72, C75.1-C75.3, D32-D33, D35.2-D35.4, D42-D43, D44.3-D44.5 | 4,487 | 387 | 519 | 67 | 5,460 | 16.9 | 24.0 | 18.9 | 7.8 | 17.2 | 8 | 3.2% |
| Breast | C50 | 44,540 | 2,840 | 4,678 | 1,294 | 53,352 | 169.8 | 175.4 | 169.3 | 154.6 | 169.7 | 1 | 30.9% |
| Breast In Situ | D05 | 6,324 | 373 | 405 | 165 | 7,267 | 24.4 | 23.7 | 14.7 | 19.4 | 23.4 | | |
| Cancer of Unknown Primary | C77-C80 | 4,021 | 306 | 460 | 115 | 4,902 | 14.6 | 17.9 | 16.5 | 13.9 | 15.0 | | 2.8% |
| Cervix | C53 | 2,639 | 146 | 318 | 104 | 3,207 | 9.8 | 9.5 | 11.6 | 11.6 | 10.0 | 12 | 1.9% |
| Cervix In Situ | D06 | 25,616 | 2,025 | 2,581 | 1,096 | 31,318 | 88.9 | 133.9 | 92.0 | 113.0 | 92.0 | | |
| Eye | C69 | 252 | 22 | 30 | 2 | 306 | 0.9 | 1.4 | 1.1 | 0.3 | 0.9 | | 0.2% |
| Gallbladder | C23 | 523 | 34 | 49 | 16 | 622 | 2.0 | 2.0 | 1.8 | 1.9 | 2.0 | | 0.4% |
| Hodgkin Lymphoma | C81 | 726 | 32 | 69 | 27 | 854 | 2.7 | 2.0 | 2.5 | 2.9 | 2.6 | | 0.5% |
| Kidney | C64-C66,C68 | 3,646 | 232 | 405 | 135 | 4,418 | 13.9 | 14.1 | 14.9 | 16.2 | 14.1 | 10 | 2.6% |
| Larynx | C32 | 290 | 29 | 67 | 14 | 400 | 1.1 | 1.8 | 2.5 | 1.7 | 1.3 | | 0.2% |
| Leukaemia | C91-C95 | 3,165 | 240 | 244 | 67 | 3,716 | 11.8 | 14.3 | 8.8 | 7.7 | 11.6 | 11 | 2.2% |
| Acute Lymphoblastic Leukaemia | C91.0 | 322 | 20 | 31 | 8 | 381 | 1.1 | 1.2 | 1.1 | 0.8 | 1.1 | | 0.2% |
| Chronic Lymphocytic Leukaemia | C91.1 | 1,098 | 91 | 71 | 18 | 1,278 | 4.2 | 5.4 | 2.6 | 2.3 | 4.0 | | 0.7% |
| Acute Myeloid Leukaemia | C92.0, C92.4, C92.5, C92.6, C92.8, C93.0, C94.0, C94.2 | 1,026 | 86 | 89 | 26 | 1,227 | 3.8 | 5.2 | 3.3 | 3.0 | 3.8 | | 0.7% |
| Chronic Myeloid Leukaemia | C92.1 | 282 | 11 | 19 | 7 | 319 | 1.1 | 0.7 | 0.7 | 0.8 | 1.0 | | 0.2% |
| Liver | C22 | 1,588 | 114 | 179 | 41 | 1,922 | 6.0 | 6.6 | 6.4 | 4.9 | 6.0 | 19 | 1.1% |
| Lung | C33-C34 | 16,823 | 1,057 | 2,625 | 539 | 21,044 | 64.4 | 64.0 | 96.1 | 66.8 | 67.2 | 2 | 12.2% |
| Malignant Melanoma | C43 | 6,213 | 347 | 590 | 207 | 7,357 | 23.5 | 21.8 | 21.4 | 24.0 | 23.2 | 5 | 4.3% |
| Mesothelioma | C45 | 387 | 17 | 37 | 4 | 445 | 1.5 | 1.0 | 1.4 | 0.5 | 1.4 | | 0.3% |
| Myeloma | C90 | 2,006 | 132 | 168 | 49 | 2,355 | 7.6 | 7.8 | 6.1 | 5.9 | 7.5 | 18 | 1.4% |
| Non-Hodgkin Lymphoma | C82-C86 | 5,206 | 293 | 513 | 142 | 6,154 | 19.9 | 18.0 | 18.9 | 17.2 | 19.7 | 7 | 3.6% |
| Oesophagus | C15 | 2,378 | 141 | 348 | 65 | 2,932 | 8.9 | 8.5 | 12.5 | 7.9 | 9.2 | 13 | 1.7% |
| Oral | C00-C06,C09-C10,C12-C14 | 2,014 | 135 | 285 | 54 | 2,488 | 7.7 | 8.4 | 10.4 | 6.4 | 8.0 | 16 | 1.4% |
| Ovary | C56-C57.4 | 6,102 | 413 | 595 | 174 | 7,284 | 23.4 | 25.5 | 21.7 | 20.7 | 23.3 | 6 | 4.2% |
| Pancreas | C25 | 3,923 | 248 | 395 | 126 | 4,692 | 14.7 | 14.8 | 14.3 | 15.7 | 14.7 | 9 | 2.7% |
| Stomach | C16 | 1,967 | 165 | 297 | 74 | 2,503 | 7.3 | 9.5 | 10.6 | 9.0 | 7.8 | 15 | 1.4% |
| Thyroid | C73 | 2,046 | 78 | 188 | 49 | 2,361 | 7.7 | 5.1 | 6.9 | 5.6 | 7.4 | 17 | 1.4% |
| Uterus | C54-C55 | 7,442 | 542 | 790 | 248 | 9,022 | 29.0 | 33.7 | 29.0 | 30.3 | 29.3 | 4 | 5.2% |
| Vagina | C52 | 179 | 20 | 29 | 8 | 236 | 0.7 | 1.2 | 1.0 | 0.9 | 0.7 | | 0.1% |
| Vulva | C51 | 1,061 | 78 | 142 | 32 | 1,313 | 4.0 | 4.8 | 5.1 | 3.8 | 4.1 | 20 | 0.8% |
| All Cancers | C00-C97 | 164,486 | 10,985 | 21,003 | 6,026 | 202,500 | 622.4 | 667.6 | 761.8 | 724.1 | 639.8 | | |
| Non-Melanoma Skin Cancer (NMSC) | C44 | 21,624 | 1,767 | 4,753 | 1,602 | 29,746 | 80.2 | 104.9 | 171.7 | 192.9 | 92.4 | | |
| **All Cancers excluding NMSC** | **C00-C97 excl C44** | **142,862** | **9,218** | **16,250** | **4,424** | **172,754** | **542.2** | **562.6** | **590.2** | **531.1** | **547.3** | | **100.0%** |
| Population estimates (2013, millions) | | 27.3 | 1.6 | 2.7 | 0.9 | 32.6 | | | | | | | |

CNS = Central Nervous System
Rates are European age-standardised and given per 100,000 population

Source: cruk.org/cancerstats



CANCER RESEARCH UK

**Figure S1: Women's Cancer statistics in 2013 from Cancer Research UK.** Reproduced from Cancer Research UK [314].

# Appendix B

# Supplementary Tables

**Table S1: Total body naevus count DMPs that passed FDR 10%.**

**Begin of multi-page Table**

| Rank | CpG | Position (hg19) | Gene | Location | CpG density | Beta | St. Error | *P* value | FDR |
|---|---|---|---|---|---|---|---|---|---|
| 1 | cg06244240 | chr17:8,1058,948 | - | - | Shore | 0.0052 | $9 \times 10^{-4}$ | $5.5 \times 10^{-8}$ | 5% |
| 2 | cg06123942 | chr15:45,722,795 | *C15orf48* | 5'UTR | Island | -0.0074 | 0.0014 | $2.2 \times 10^{-7}$ | 5% |
| 3 | cg25384157 | chr9:140,499,131 | *ARRDC1* | TSS 1500 | Shore | 0.0063 | 0.0012 | $3.1 \times 10^{-7}$ | 5% |
| 4 | cg11297934 | chr3:12,705,868 | *RAF1* | TSS 200 | Island | -0.0046 | $9 \times 10^{-4}$ | $1.2 \times 10^{-6}$ | 10% |
| 5 | cg14762973 | chr2:187,714,067 | *ZSWIM2* | TSS 200 | Island | 0.0069 | 0.0014 | $1.49 \times 10^{-6}$ | 10% |
| 6 | cg15977816 | chr19:44,285,297 | *KCNN4* | 5'UTR; 1st Exon | - | 0.0061 | 0.0013 | $2.2 \times 10^{-6}$ | 10% |
| 7 | cg03755177 | chr7:44,349,389 | *CAMK2B* | Body | Island | -0.0066 | 0.0014 | $2.3 \times 10^{-6}$ | 10% |
| 8 | cg01880437 | chr4:26,321,873 | *RBPJ* | TSS 1500 | Island | -0.0057 | 0.0012 | $3.1 \times 10^{-6}$ | 10% |
| 9 | cg02683509 | chr6:156,950,855 | - | - | Shore | 0.0059 | 0.0012 | $3.2 \times 10^{-6}$ | 10% |
| 10 | cg18401367 | chr5:176,107,201 | - | - | Island | -0.0063 | 0.0013 | $3.3 \times 10^{-6}$ | 10% |
| 11 | cg11074933 | chr15:26,915,414 | *GABRB3* | Body | Island | -0.0063 | 0.0013 | $3.7 \times 10^{-6}$ | 10% |
| 12 | cg03163343 | chr1:45,118,395 | - | - | Shore | 0.0049 | 0.001 | $4.7 \times 10^{-6}$ | 10% |
| 13 | cg20093198 | chr19:24,182,837 | - | - | - | 0.0053 | 0.0011 | $5.1 \times 10^{-6}$ | 10% |

**Continuation of Table on next page**

**Continuation of Table S1**

| Rank | CpG | Position (hg19) | Gene | Location | CpG density | Beta | St. Error | *P* value | FDR |
|---|---|---|---|---|---|---|---|---|---|
| 14 | cg06569947 | chr13:114,126,889 | *DCUN1D2* | Body | - | -0.006 | 0.0013 | 5.3 x $10^{-6}$ | 10% |
| 15 | cg09536336 | chr5:168,439,657 | *SLIT3* | Body | - | 0.0053 | 0.0012 | 5.7 x $10^{-6}$ | 10% |
| 15 | cg09536336 | chr5:168,439,657 | *SLIT3* | Body | - | 0.0053 | 0.0012 | 5.7 x $10^{-6}$ | 10% |
| 16 | cg25343280 | chr2:44,314,198 | - | - | Shore | 0.0055 | 0.0012 | 6.1 x $10^{-6}$ | 10% |
| 17 | cg21068480 | chr2:85,980,500 | *ATOH8* | TSS 1500 | Island | 0.0057 | 0.0012 | 6.5 x $10^{-6}$ | 10% |
| 18 | cg22534759 | chr10:32,403,029 | - | - | Shelf | 0.0044 | 0.001 | 6.5 x $10^{-6}$ | 10% |
| 19 | cg00347643 | chr7:75,957,202 | *YWHAG* | 3'UTR | Shore | -0.0053 | 0.0011 | 6.8 x $10^{-6}$ | 10% |
| 20 | cg02236913 | chr1:20,005,598 | *HTR6* | Body | Island | 0.0059 | 0.0013 | 6.8 x $10^{-6}$ | 10% |
| 21 | cg25720825 | chr1:2,849,682 | - | - | Shore | 0.0053 | 0.0012 | 7.5 x $10^{-6}$ | 10% |
| 22 | cg23082845 | chr1:1,159,282 | *SDF4* | Body | Island | 0.0056 | 0.0012 | 7.8 x $10^{-6}$ | 10% |
| 23 | cg14278345 | chr22:50,451,054 | *IL17REL* | 5'UTR; 1st Exon | Shore | 0.0046 | 0.001 | 7.9 x $10^{-6}$ | 10% |
| 24 | cg10139717 | chr8:2,363,092 | - | - | - | -0.0056 | 0.0012 | 8.1 x $10^{-6}$ | 10% |
| 25 | cg10929758 | chr2:54,857,270 | *SPTBN1* | Body | Shore | 0.0051 | 0.0011 | 8.2 x $10^{-6}$ | 10% |
| 26 | cg06739855 | chr1:17,240,204 | - | - | Island | -0.0057 | 0.0013 | 8.5 x $10^{-6}$ | 10% |

**Continuation of Table on next page**

**Continuation of Table S1**

| Rank | CpG | Position (hg19) | Gene | Location | CpG density | Beta | St. Error | *P* value | FDR |
|---|---|---|---|---|---|---|---|---|---|
| 27 | cg02496234 | chr7:20,826,128 | *SP8* | 5'UTR | Shore | 0.0061 | 0.0013 | $8.8 \times 10^{-6}$ | 10% |
| 28 | cg09456289 | chr2:43,038,399 | - | - | Island | -0.0063 | 0.0014 | $9.4 \times 10^{-6}$ | 10% |
| 29 | cg03315230 | chr2:27,346,938 | *ABHD1* | Body | Shore | 0.0045 | 0.001 | $9.6 \times 10^{-6}$ | 10% |
| 30 | cg07821574 | chr19:18,208,505 | *MAST3* | TSS 200 | Shore | -0.0042 | $9 \times 10^{-4}$ | $9.7 \times 10^{-6}$ | 10% |
| 31 | cg22985929 | chr19:6,227,536 | *MLLT1* | Body | Shelf | 0.005 | 0.0011 | $9.7 \times 10^{-6}$ | 10% |
| 32 | cg04685954 | chr12:10,515,574 | - | - | - | 0.0043 | 0.001 | $9.8 \times 10^{-6}$ | 10% |
| 33 | cg20193802 | chr22:19709696 | *SEPT5*; *GP1BB* | Body; TSS 1500 | Island | 0.005 | 0.0011 | $1.0 \times 10^{-5}$ | 10% |
| 34 | cg21220374 | chr9:100,396,777 | *TSTD2*; *NCBP1* | TSS 1500; Body | Shore | 0.0058 | 0.0013 | $1.1 \times 10^{-5}$ | 10% |
| 35 | cg19179973 | chr6:83,775,554 | *UBE2CBP* | TSS 200 | Island | 0.004 | $9 \times 10^{-4}$ | $1.1 \times 10^{-5}$ | 10% |
| 36 | cg04587829 | chr17:80,692,947 | *FN3K* | TSS 1500 | Shore | 0.0048 | 0.0011 | $1.1 \times 10^{-5}$ | 10% |
| 37 | cg01505254 | chr6:36,853,867 | *C6orf89* | Body | Island | -0.0053 | 0.0012 | $1.2 \times 10^{-5}$ | 10% |
| 38 | cg25369015 | chr1:9,599,256 | *SLC25A33* | TSS 1500 | Shore | -0.0063 | 0.0014 | $1.3 \times 10^{-5}$ | 10% |

**Continuation of Table on next page**

**Continuation of Table S1**

| Rank | CpG | Position (hg19) | Gene | Location | CpG density | Beta | St. Error | *P* value | FDR |
|---|---|---|---|---|---|---|---|---|---|
| 39 | cg22494876 | chr1:247,618,161 | - | - | Shelf | 0.006 | 0.0014 | 1.3 x $10^{-5}$ | 10% |
| 40 | cg13631572 | chr14:24,803,903 | *ADCY4* | TSS 200 | Island | -0.0058 | 0.0013 | 1.4 x $10^{-5}$ | 10% |
| 41 | cg03166286 | chr9:88,555,512 | *NAA35* | TSS 1500 | Shore | 0.0058 | 0.0013 | 1.4 x $10^{-5}$ | 10% |
| 42 | cg15375596 | chr15:99,434,736 | *IGF1R* | Body | Island | 0.0053 | 0.0012 | 1.4 x $10^{-5}$ | 10% |
| 43 | cg22617002 | chr13:112,275,633 | - | - | Island | 0.0054 | 0.0012 | 1.5 x $10^{-5}$ | 10% |
| 44 | cg18421356 | chr3:134,204,114 | *ANAPC13*; *CEP63* | 5'UTR; TSS 1500 | Shore | -0.005 | 0.0011 | 1.5 x $10^{-5}$ | 10% |
| 45 | cg13650689 | chr19:54,618,596 | *PRPF31*; *TFPT* | TSS 200; Body | Shore | -0.0049 | 0.0011 | 1.5 x $10^{-5}$ | 10% |
| 46 | cg10339152 | chr20:749,452 | *C20orf54* | TSS 1500 | - | 0.0041 | 9 x $10^{-4}$ | 1.5 x $10^{-5}$ | 10% |
| 47 | cg02208506 | chr17:4,634,804 | *MED11* | 1st Exon | Island | -0.0055 | 0.0013 | 1.6 x $10^{-5}$ | 10% |
| 48 | cg12836303 | chr11:63,756,920 | *OTUB1* | Body | Shelf | 0.0059 | 0.0013 | 1.6 x $10^{-5}$ | 10% |

**End of Table**

**Table S2: Total body naevus count DMRs with $p$ value <0.01.**

**Begin of multi-page Table**

| Rank | Position (hg19) | Gene | Location | CpG density | Number of CpGs | Direction | $P$ value | Direction CpG sites |
|---|---|---|---|---|---|---|---|---|
| 1 | chr9:140,499,132-140,500,813 | *ARRDC1* | TSS 1500 - Body | Island | 7 | + | 2.6 x $10^{-5}$ | ++ - - - ++ |
| 2 | chr10:14,647,154-14,647,530 | *FAM107B* | Body | Shore | 3 | + | 2.5 x $10^{-4}$ | +++ |
| 3 | chr19:44,285,297-44,285,568 | *KCNN4* | TSS 200 - 1st Exon | - | 3 | + | 2.9 x $10^{-4}$ | +++ |
| 4 | chr17:8,129,997-8,130,356 | *CTC1* | 3'UTR | Shelf | 3 | - | 6.3 x $10^{-4}$ | - - - |
| 5 | chr15:26,915,414-26,915,752 | *GABRB3* | Body | Island | 3 | - | 8.3 x $10^{-4}$ | - - - |
| 6 | chr1:165,513,318-165,513,343 | *LOC400794*; *LRRC52* | Body; TSS 200 | - | 3 | + | 9.6 x $10^{-4}$ | +++ |
| 7 | chr7:44,349,389-44,349,389 | *CAMK2B* | Body | Island | 1 | - | 0.0012 | - |
| 8 | chr8:130,995,990-130,996,123 | - | - | Island | 3 | - | 0.0012 | - - - |
| 9 | chr13:114,126,889-114,126,889 | *DCUN1D2* | Body | - | 1 | - | 0.0013 | - |
| 10 | chr11:64,739,320-64,739,343 | - | - | Island | 3 | - | 0.0018 | - - - |
| 11 | chr10:102,279,455-102,279,694 | *SEC31B* | TSS 200 - 5'UTR | Island | 3 | - | 0.0022 | - - - |
| 12 | chr19:43,968,133-43,968,495 | *LYPD3* | Body | Island | 2 | - | 0.0022 | - - |
| 13 | chr2:4,600,947-4,601,053 | - | - | - | 3 | + | 0.0025 | +++ |

**Continuation of Table on next page**

**Continuation of Table S2**

| Rank | Position (hg19) | Gene | Location | CpG density | Number of CpGs | Direction | *P* value | Direction CpG sites |
|---|---|---|---|---|---|---|---|---|
| 14 | chr20:5,485,270-5,485,294 | *LOC149837* | TSS 200 | - | 3 | - | 0.0026 | - - - |
| 15 | chr7:152,063,901-152,063,974 | *MLL3* | Body | Island | 3 | - | 0.0026 | - - - |
| 16 | chr7:98,030,324-98,030,641 | *BAIAP2L1* | TSS 1500 - TSS 200 | Shore | 4 | + | 0.0028 | ++ - + |
| 17 | chr1:151,693,222-151,693,261 | *C1orf230* | TSS 1500 | Shore | 2 | - | 0.0028 | - - |
| 18 | chr1:17,240,204-17,240,204 | - | - | Island | 1 | - | 0.0029 | - |
| 19 | chr8:82,633,130-82,633,568 | *ZFAND1* | TSS 200 - Body | Island | 4 | - | 0.003 | - + - - |
| 20 | chr11:102,576,469-102,576,508 | *MMP27* | TSS 200 | - | 2 | + | 0.0034 | ++ |
| 21 | chr2:85,980,500-85,980,500 | *ATOH8* | TSS 1500 | Island | 1 | + | 0.0034 | + |
| 22 | chr1:27,729,801-27,729,992 | - | - | - | 2 | + | 0.0035 | ++ |
| 23 | chr14:38,091,400-38,091,470 | - | - | Shore | 2 | - | 0.0038 | - - |
| 24 | chr4:26,321,873-26,321,873 | *RBPJ* | TSS 1500 | Island | 1 | - | 0.0038 | - |
| 25 | chr12:1,905,735-1,906,097 | *CACNA2D4* | Body | Island | 2 | + | 0.0038 | ++ |
| 26 | chr18:56,296,449-56,296,607 | *ALPK2* | TSS 1500 | - | 4 | + | 0.0039 | ++++ |
| 27 | chr22:45,598,944-45,599,059 | *C22orf9* | Body | Island | 2 | + | 0.0039 | ++ |

**Continuation of Table on next page**

**Continuation of Table S2**

| Rank | Position (hg19) | Gene | Location | CpG density | Number of CpGs | Direction | *P* value | Direction CpG sites |
|---|---|---|---|---|---|---|---|---|
| 28 | chr1:9,599,256-9,599,276 | *SLC25A33* | TSS 1500 | Shore | 2 | - | 0.0042 | - - |
| 29 | chr7:105,331,648-105,331,690 | *ATXN7L1* | Body | - | 2 | - | 0.0042 | - - |
| 30 | chr16:2,502,146-2,502,146 | *CCNF* | Body | - | 1 | - | 0.0044 | - |
| 31 | chr1:1,159,282-1,159,282 | *SDF4* | Body | Island | 1 | + | 0.0048 | + |
| 32 | chr1:43,770,557-43,770,707 | *TIE1* | Body | Island | 2 | - | 0.0051 | - - |
| 33 | chr17:4,634,804-4,634,804 | *MED11* | 1st Exon | Island | 1 | - | 0.0055 | - |
| 34 | chr1:162,039,224-162,039,224 | *NOS1AP* | TSS 1500 | Shore | 1 | - | 0.0056 | - |
| 35 | chr14:6,5016,591-65,016,602 | *C14orf50* | TSS 200 | Island | 3 | + | 0.0059 | +++ |
| 36 | chr12:54,132,255-54,132,568 | - | - | Shore | 3 | - | 0.0064 | - - - |
| 37 | chr17:6,347,533-6,347,533 | *FAM64A* | TSS 1500 | Island | 1 | - | 0.0067 | - |
| 38 | chr2:44,314,198-44,314,198 | - | - | Shore | 1 | + | 0.0069 | + |
| 39 | chr16:28,565,199-28,565,206 | *CCDC101* | TSS 200 | Island | 2 | - | 0.0072 | - - |
| 40 | chr4:55,092,375-55,092,556 | - | - | Shore | 2 | + | 0.0074 | ++ |

**Continuation of Table on next page**

**Continuation of Table S2**

| Rank | Position (hg19) | Gene | Location | CpG density | Number of CpGs | Direction | *P* value | Direction CpG sites |
|---|---|---|---|---|---|---|---|---|
| 41 | chr7:152,456,579-152,456,597 | *ACTR3B* | TSS 1500 | Shore | 2 | + | 0.0082 | ++ |
| 42 | chr22:39,712,694-39,712,730 | *RPL3*; | Body; | Shelf | 2 | + | 0.0084 | ++ |
| | | *SNORD83A* | TSS 1500 | | | | | ++ |
| 43 | chr9:100,396,777-100,396,777 | *TSTD2*; | TSS 1500; | Shore | 1 | + | 0.0089 | + |
| | | *NCBP1* | Body | | | | | |
| 44 | chr17:78,865,087-78,866,235 | *RPTOR* | Body | Shore | 8 | + | 0.0089 | ++++++++ |
| 45 | chr18:43,548,144-43,548,144 | *KIAA1632* | TSS 1500 | Shore | 1 | + | 0.0091 | + |
| 46 | chr12:52,695,412-52,695,515 | *KRT86* | TSS 200 | Shore | 2 | - | 0.0094 | - - |
| 47 | chr2:3,471,345-3,471,345 | *TTC15* | Body | - | 1 | + | 0.0097 | + |
| 48 | chr6:31,743,928-31,743,952 | *C6orf27* | Body | - | 2 | + | 0.0098 | ++ |

**End of Table**